ICN 2011

The Tenth International Conference on Networks

January 23-28, 2011 - St. Maarten,

The Netherlands Antilles

ICN 2011 Editors

Pascal Lorenz, University of Haute Alsace, France

Tibor Gyires, Illinois State University, USA

Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland

ICN 2011

Foreword

The Tenth International Conference on Networking (ICN 2011), held on January 23-27, 2011 in St. Maarten, The Netherlands Antilles, addressed hot topics in networking.

ICN 2011 gathered international scientists - researchers, practitioners, and students - interested in new developments targeting all areas of networking. The accepted papers covered a wide range of networking-related topics spanning from protocols, architectures, P2P, network performance, QoS/reliability, signal processing, communications theory, security, multimedia/multicast, network management and control, vehicular networks, and wireless networks. We believe that the ICN 2011 contributions offered a large panel of solutions to key problems in all areas of global knowledge and set challenging directions for industrial research and development.

We take this opportunity to thank all the members of the ICN 2011 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to the ICN 2011. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICN 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICN 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in networking research.

The beautiful places of St. Maarten surely provided a pleasant environment during the conference and we hope you had a chance to visit the surroundings.

ICN 2011 Chairs

Tibor Gyires, Illinois State University, USA
Eva Hladká, Masaryk University - Brno / CESNET, Czech Republic
Pascal Lorenz, University of Haute Alsace, France
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION, Spain
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France

ICN 2011

Committee

ICN Advisory Chairs

Tibor Gyires, Illinois State University, USA
Eva Hladká, Masaryk University - Brno / CESNET, Czech Republic
Pascal Lorenz, University of Haute Alsace, France
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION, Spain
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France

ICN 2011 Technical Program Committee

Kari Aho, University of Jyväskylä, Finland
Khaled Amleh, Penn State University - Mont Alto, USA
Cristian Anghel, Politehnica University of Bucharest, Romania
Tarun Bansal, The Ohio State University - Columbus, USA
Alvaro Barradas, University of Algarve, Portugal
Zdenek Becvar, Czech Technical University in Prague, Czech Republic
Thomas Beluch, Université de Toulouse, France
Jalel Ben-Othman, Université de Versailles, France
Djamel Benferhat, University Of South Brittany, France
Robert Bestak, Czech Technical University in Prague, Czech Republic
Jun Bi, Tsinghua University, China
Frank Bohdanowicz, University of Koblenz, Germany
Jean-Marie Bonnin, Institut Télécom / Télécom Bretagne, France
Fernando Boronat Seguí, Universidad Politécnica de Valencia, Spain
Detlef Bosau, Scientist-Stuttgart, Germany
Matthias R. Brust, Technological Institute of Aeronautics, Brazil
Alexandre Caminada, Laboratoire Systèmes et Transport (UTBM) - Belfort, France
Eduardo Cerqueira, Federal University of Para, Brazil
Marc Cheboldaeff, Alcatel-Lucent Deutschland AG - Ratingen, Germany
Kwangjong Cho, Korea Institute of Science and Technology Information (KISTI) - Daejeon, Republic of Korea
Andrzej Chydzinski, Silesian University of Technology - Gliwice, Poland
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION, Spain
Weibei Dou, Tsinghua University-Beijing, P.R.China
Daniela Dragomirescu, LAAS/CNRS, Toulouse, France
Gledson Elias, Federal University of Paraíba, Brazil
Khalid Farhan, Al-Zaytoonah University of Jordan, Jordan
Mário F. S. Ferreira, University of Aveiro, Portugal
Luciana Andreia Fondazzi Martimiano, Universidade Estadual de Maringá, Brazil
Mário Freire, University of Beira Interior, Portugal
Wolfgang Fritz, Leibniz Supercomputing Centre - Garching b. München, Germany
Holger Fröning, University of Heidelberg, Germany

Laurent George, University of Paris-Est Creteil Val de Marne, France
Eva Gescheidtova, Brno University of Technology, Czech Republic
Markus Goldstein, German Research Center for Artificial Intelligence (DFKI), Germany
Anahita Gouya, AFD Technologies, France
Vic Grout, Glyndwr University - Wrexham, UK
Mina S. Guirguis, Texas State University - San Marcos, USA
Huaqun Guo, Institute for Infocomm Research, A*STAR, Singapore
Tibor Gyires, Illinois State University, USA
Keijo Haataja, University of Eastern Finland- Kuopio / Unicta Oy, Finland
Timo Hämäläinen, University of Jyvaskyla, Finland
Mohammad Hammoudeh, Manchester Metropolitan University, UK
Oliver Hanka, TU München, Germany
Eva Hladka, Masaryk University - Brno / CESNET, Czech Republic
Chen-Shie Ho, Oriental Institute of Technology, Taiwan
Osamu Honda, Onomichi University, Japan
Florian Huc, University of Geneva, Switzerland
Jin-Ok Hwang, Korea University - Seoul, Korea
Muhammad Ali Imran, University of Surrey - Guildford, UK
Norbert Jordan, Accenture, Austria
Omid Kashefi, Iran University of Science and Technology-Tehran, Iran
Andrzej Kasprzak, Wroclaw University of Technology, Poland
Kazuhiko Kinoshita, Osaka University, Japan
Joséphine Kohlenberg, TELECOM SudParis, France
Markku Kojo, University of Helsinki, Finland
Leszek Koszalka, Wroclaw University of Technology, Poland
Tomas Koutny, University of West Bohemi-Pilsen, Czech Republic
Polychronis Koutsakis, Technical University of Crete, Greece
Hadi Larijani, Glasgow Caledonian University, UK
Angelos Lazaris, University of Southern California, USA
Steven S. W. Lee, National Chung Cheng University, Taiwan R.O.C.
Lada-On Lertsuwanakul, FernUniversität - Hagen, Germany
Xining Li, University of Guelph, Canada
Miloš Liška, CESNET / Masaryk University - Brno, Czech Republic
Pascal Lorenz, University of Haute Alsace, France
Pavel Mach, Czech Technical University in Prague, Czech Republic
Damien Magoni, University of Bordeaux, France
Ahmed Mahdy, Texas A&M University - Corpus Christi, USA
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Rui Marinheiro, ISCTE - Lisbon University Institute, Portugal
Teo Yong Meng, National University of Singapore / Sun Microsystems Inc., Singapore
Pascale Minet, INRIA-Rocquencourt, France
Jogesh Muppala , Hong Kong University of Science and Technology, Hong Kong
Michael Müter, Daimler AG - Böblingen, Germany
Katsuhiro Naito, Mie University - Tsu City, Japan
Frank Oldewurtel, RWTH Aachen University, Germany
Go-Hasegawa, Osaka University, Japan
Jose Oscar Fajardo, University of the Basque Country - Bilbao, Spain
Constantin Paleologu, University Politehnica of Bucharest, Romania

Nikolaos A. Pantazis, Technological Educational Institution (TEI) of Athens, Greece
Konstantinos Patsakis, University of Piraeus, Greece
Ionut Pirnog, Politehnica University of Bucharest, Romania
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland
Jani Puttonen, Magister Solutions Ltd., Finland
Yingzhen Qu, Cisco Systems, Inc., USA
Milena Radenkovic, University of Nottingham, UK
Victor Ramos, UAM-Iztapalapa, Mexico
Priyanka Rawat, Telecom SudParis, France
Krisakorn Rerkrai, RWTH Aachen University, Germany
Karim Mohammed Rezaul, Glyndwr University - Wrexham, UK
Joel Rodrigues, Intituto de Telecomunicações / University of Beira Interior, Portugal
Wouter Rogiest, Ghent University, Belgium
Teerapat Sanguankotchakorn, Asian Institute of Technology - Klong Luang, Thailand
Rajarshi Sanyal, Belgacom-ICS, Belgium
Susana Sargento, University of Aveiro, Portugal
Masahiro Sasabe, Osaka University, Japan
Raimund Schatz, FTW Forschungszentrum Telekommunikation Wien GmbH, Austria
Thomas C. Schmidt, HAW Hamburg, Germany
Hans Scholten, University of Twente- Enschede, The Netherlands
Lijie Sheng, Xidian University - Xi'an, China
Karel Slavicek, Masaryk University Brno, Czech Republic
Lars Strand, Norwegian Computing Center, Norway
Miroslav Sveda, Brno University of Technology, Czech Republic
Nabil Tabbane, SUPCOM- Higher National School of Telecommunications -Tunis, Tunisia
Ken Turner, The University of Stirling, UK
Manos Varvarigos, University of Patras, Greece
Dario Vieira, EFREI, France
Lukas Vojtech, Czech Technical University in Prague, Czech Republic
Joris Walraevens, SMACS / Ghent University - Ugent, Belgium
Gary Weckman, Ohio University, USA
Maarten Wijnants, Hasselt University-Diepenbeek, Belgium
Qin Xin, Simula Research Laboratory - Oslo, Norway
Qimin Yang, Harvey Mudd College-Claremont, USA
Vladimir Zaborovski, Polytechnic University of Saint Petersburg, Russia
Arkady Zaslavsky, Luleå University of Technology, Sweden
Hans-Jürgen Zepernick, Blekinge Institute of Technology, Sweden

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

Trust Model-based Secure Cooperative Sensing Techniques for Cognitive Radio Networks

Deming Pang, Gang Hu, Ming Xu
School of Computer, National University of Defense Technology
Changsha, China
e-mail: pang3724@nudt.edu.cn, golfhg@vip.sohu.net, xuming-64@hotmail.com

***Abstract*—Cooperative spectrum sensing has been shown to enable Cognitive Radio (CR) networks to reliably detect licensed users and avoid causing interference to licensed communications. However, the performance of the scheme can be severely degraded due to presence of malicious users sending false sensing data. In this paper, we propose trust model-based cooperative sensing techniques to reduce the harmful effect of malicious users in cooperative sensing process. First of all, analysis model of anomalous behavior is devised to identify the malicious users. Then we employ PID (Proportional-Integral-Derivative) like controller to calculate the credit value of nodes, which is used as the weight of WBD (Weighted Bayesian Detection) to make spectrum decision. Simulation results demonstrate that, comparing with the existing methods, the proposed scheme performs better especially in the case that there exist a large number of malicious nodes.**

Keywords- Cognitive Radio Networks; Cooperative Spectrum Sensing; Trust Model; Weighted Bayesian Detection.

I. INTRODUCTION

Cognitive radio (CR) techniques provide the capability to use or share the spectrum in an opportunistic manner, which is proposed to solve current spectrum inefficiency problem [1]. In CR networks unlicensed users (secondary users, SUs) detect spectrum environment and utilize the idle spectrum while tolerable interference is guaranteed to the licensed users (primary users, PUs).

For CR networks, reliable spectrum sensing is an important step for any practical deployment. SUs should identify the presence of PUs over wide range of spectrum accurately without significant delay. This process is very difficult as we need to identify various PUs adopting different modulation schemes, data rates and transmission powers in presence of variable propagation losses, interference generated by other secondary users and thermal noise. Traditionally there are three spectrum sensing techniques, viz., energy detection, matched filter detection and cyclostationary feature detection [2]. If SUs are lack of knowledge about the characteristics of PU signal, energy detection is the optimal choice with the least complexity and generally adopted in recent research work. However, the performance of energy detection is always degraded because of signal-to-noise ratio wall or channel fading/shadowing [3].

Cooperation among SUs follows almost as a necessary consequence of the above constraints. Cooperative spectrum sensing has been shown to greatly increase the probability of detecting the PUs [4-6]. Each SU executes spectrum sensing by itself and sends the "local" spectrum sensing information to a DC (Data Collector) which uses an appropriate data fusion technique to make final spectrum sensing decision.

Since a DC utilises not only its own observations as a basis for decision making but also the observations of others, it is the obviously need to authenticate the shared observations. The DC needs to judge whether the observations from others are real or falsified. This is critical to prevent degradation of the network performance because of malicious behavior and to protect against the Byzantine attack. The Byzantine attack represents the case where a friend or acquaintance has, unbeknownst to the CR, become an adversary and represents the most difficult subset of this problem space. The Byzantine failure problem can be caused by malfunctioning sensing terminals or MUs (Malicious Users). They transmit false information instead of real detection results, which adversely affects the global decision.

This problem has been discussed in [7-9], and several methods were proposed to reduce the impact of false information. But these proposals failed when the proportion of MUs increased. In this paper, we investigate techniques to identify the nodes which provide false sensing information, and nullify their effect on the cooperative spectrum sensing system. By analyzing the behaviors of SUs in cooperative sensing, DC can establish trust model with PID (Proportional-Integral-Derivative) like controller [10], which can acquire relatively high speed to track the behaviors of neighbors. At last, we use a fusion technique called WBD (Weighted Bayesian Detection) derived from Bayesian detection to make spectrum decision. Simulation results show that this method improves the robustness of data fusion against attacks even when a large proportion of malicious users exist.

In Section II, we define the system model. In Section III, the proposed cooperative detection scheme is described in detail. Simulation results and analysis are illustrated in Section IV, and finally, a conclusion is given in Section V.

II. SYSTEM MODEL

In an ad hoc CR network, we consider a group of N second users in the presence of a primary user working on K different channels. The channels of PU and SUs use the HATA model for rural environments as the path loss model [11]. We assume perfect channel conditions for the control channel. Each of the SUs acts as a sensing terminal that is responsible for local spectrum sensing. The local detection results are reported to a DC that executes data fusion and makes the final spectrum decision. SUs use energy detector,

 ISBN:978-1-61208-113-7

and the sensing report is local sensing decision which is a binary variable— "1" denotes the presence of PU signal, and "0" denotes its absence. The data fusion problem therefore can be regarded as a binary hypothesis testing problem with two hypotheses represented by H1 and H0 (H1 means there exists primary user, and H1 means the channel is free). We consider three types of spectrum spoofing attacks: always-false, always-busy and always-free. An always-false attacker always sends spectrum reports that are opposite to its real local sensing results, and an always-busy attacker always notifies spectrum to be busy while an always-free attacker always reports contrary results.

III. TRUST MODEL-BASED SECURE COOPERATIVE SENSING

This Section will detail a reactive protection mechanism, a trust model-based cooperation enforcement mechanism to improve robustness of data fusion technique. First of all, the anomalous behaviors of malicious users should be identified. We design two kinds of behavior analysis models to track the behaviors of SUs. Based on that observation, a sensing terminal's reputation can be calculated with a PID-like trust model. Since the data fusion is a binary hypothesis testing problem, we propose a new technique called WBD to overcome the weakness of existing fusion techniques.

A. Analysis of Anomalous Behaviors

After receiving local reports of neighbors, DC should judge which one is believable, and make spectrum decision with appropriate reports. DC will get the ultimate sensing result *U* at the end of sensing period, which derives from *i* neighbors sensing *k* channels.

$$U=\begin{pmatrix} c_{11} & \cdots & c_{1k} \\ \vdots & \ddots & \vdots \\ c_{i1} & \cdots & c_{ik} \end{pmatrix}$$

where U is a $i\times k$ matrix which consists of 0 and 1, and c_{ik} is the sensing result of channel k detected by node i. $c_{ik}=1$ means there exists PU in channel k, and $c_{ik}=0$ means that channel k is free.

First of all, we focus on the analysis of abnormal sensing behaviors in single channel. $(c_{1j},\cdots,c_{mj},\cdots,c_{ij})$ is the sensing result of channel j. Without loss of generality, we assume the first m items are same, If $m>i/2$, m nodes correspond to these items are judged to be normal while the others are malicious. Different kinds of users will be assigned corresponding credit values with the following algorithm in Section B.

This is a kind of majority rule, which is feasible when the proportion of MUs is small. We consider CR networks with N SUs, among which M SUs are malicious. The false detection ratio with energy detection is α. The analysis of anomalous behaviors is effective under the condition

$$\begin{aligned} &\frac{(N-M)\cdot\alpha+M\cdot(1-\alpha)}{N}<50\% \\ &M<N/2 \end{aligned} \tag{1}$$

Since the correct detection ratio of energy detection is not ideal, the credit values of SUs calculated in single channel may not be assigned rightly (normal SU is regard as malicious node, whose credit value is decreased. v.v.). But in a sensing period multiple channels would be detected in the same way, and the credit value of each node will be updated in each channel. So the probability of miscalculation of credit value P_f could be shown as

$$\begin{aligned} P_f &= \sum_{i=\frac{k}{2}+1}^{k} P(a=i) \\ &= \sum_{i=\frac{k}{2}+1}^{k} C_k^i\cdot Q^i\cdot(1-Q)^{k-i} \end{aligned} \tag{2}$$

where a is the number of channels on which there exist trust misjudgment, and $P(a=i)$ means the probability that there exist i channels on which misjudgment is present. Q is the probability of misjudgment in single channel.

$$Q=P[H_1]\cdot P[0\,|\,H_1]+P[H_0]\cdot P[1\,|\,H_0] \tag{3}$$

Based on (2) (3), P_f would be very small when the regulation of evaluating sensing nodes' behavior is available, viz. the number of MUs is not more than SUs. (e.g., $k=20$, $Q=0.3$, we can deduce $P_f\approx 0.016$). So we can make a conclusion that the majority rule at multi-channel environment is effective in identifying malicious users.

When SUs have detected k channels in distributed manner in a sensing period, DC can deduce the numbers of available channels $M=(n_1,n_2,\cdots,n_i)$ from i different SUs where n_i is the number of available channels from the sensing report of node i. These numbers should be identical in theory, while difference always exists because of malfunction or intention of sensing nodes. For example, the number of available channels from the report of an always-busy attacker would be zero. In order to identify malicious behaviors by analyzing the numbers of available channels, M should be amended using prior probability to eliminate the infection of malfunction. We can make use of a proposal devised in paper [7] to get the prior probability in different channels.

$$\begin{aligned} P_{11} &= \begin{pmatrix} p_{1,1}^1 & \cdots & p_{i,1}^1 \\ \vdots & \ddots & \vdots \\ p_{1,k}^1 & \cdots & p_{i,k}^1 \end{pmatrix} = \left(P_1^1,P_2^1,\cdots,P_i^1\right) \\ P_{01} &= \begin{pmatrix} p_{1,1}^0 & \cdots & p_{i,1}^0 \\ \vdots & \ddots & \vdots \\ p_{1,k}^0 & \cdots & p_{i,k}^0 \end{pmatrix} = \left(P_1^0,P_2^0,\cdots,P_i^0\right) \end{aligned} \tag{4}$$

where $p_{i,k}^1$, $p_{i,k}^0$ means the prior probability $P[1|H_1]$ and $P[0|H_1]$ of node *i* in channel *k* respectively. Thus DC can revise the number of available channels from each SU as follows

$$\begin{aligned} U &= (C_1,\cdots,C_i)^t \\ E-U &= (D_1,\cdots,D_i)^t \\ M &= (n_1,n_2,\cdots n_i) \end{aligned} \quad (5)$$

$$= \left(C_1\cdot P_1^1 + D_1\cdot P_1^0,\cdots,C_i\cdot P_i^1 + D_i\cdot P_i^0\right) \quad (6)$$

where E is a $i\times k$ matrix which consists of 1. Through comparing the number of available channels from any node *i* with DC *j*, we can distinguish malicious nodes among SUs as follows

$$\Delta n = \left|n_j - n_i\right|/k$$

$$\begin{cases} \Delta n > \beta, & \textit{node i behaviors anormally} \\ \Delta n < \beta, & \textit{node i behaviors normally} \end{cases} \quad (7)$$

where β is a threshold to distinguish the status of nodes.

B. PID-Like Trust Model

Considering that the sensing reports influence the allocation and accessing of spectrum resource directly, the credit values should track the behaviors of SUs rapidly in order to reduce the negative influence. On the other hand, energy detection is not ideal. Mistaken sensing reports may be send to DC by normal SUs which would be seen as malicious users. So the credit values should be modified in a smooth manner in order to avoid random mistake of normal SUs. We use a tuned PID controller in control systems to calculate the trust values of nodes [10]:

$$\begin{aligned} f(t) &= B(t) - V(t) \\ V(t) &= \alpha * \int_0^t f(t)dt \end{aligned} \quad (8)$$

Under the conditions:

$$B(t)\in\{0,1\}$$
$$V(t)\in[0,1],\ V(0)=1$$

In equation (8) $B(t)$ is an input which is the detected result of the neighbor's behavior, and $V(t)$ is the corresponding output. The right of the lower equation refers to the record of history about difference. According to the control theory, the input is zero order signal, and the controller can track the input in time and get no static difference. So this model can trace the behaviors of the neighbors at a higher speed and attain a smooth change of trust value. Based on (8) we can deduce

$$V(t) = e^{-\alpha\cdot t}, \quad t\ge 0 \quad (9)$$

Equation (9) can be used to calculate the parameter α .

In order to utilize this trust model, equation (8) is discretized as a discrete equation which is shown in Equation (10).

$$\begin{aligned} f_i^j(k+1) &= B_i^j(k+1) - V_i^j(k) \\ V_i^j(k+1) &= \alpha * \sum_{n=0}^{k+1} f_i^j(n) \\ \alpha * f_i^j(0) &= 1 \end{aligned} \quad (10)$$

where $V_i^j(k+1)$ denotes the trust value of node *i* in sensing period $k+1$ recorded in DC *j*. The others have the similar meanings corresponding to those in equation (8).

$$B_i^j(k) = \begin{cases} 0, & \textit{If node i behaviors anormally at period k} \\ 1, & \textit{If node i behaviors normally at period k} \end{cases}$$

In addition, we define a threshold of trust value V_T to indicate whether DC should believe its neighbors. Using V_T and *N* to substitute *V*(*t*) and *t* in equation (9) respectively, we can deduce the parameter α as

$$\alpha = -\frac{\ln V_T}{N} \quad (11)$$

where *N* is the number of steps in which the trust value changes from 1 to V_T when the input is always 0 after certain time point, defined as the speed to trace the behaviors of neighbors. *N* can be figured out approximately as

$$P^N < P_{toler} \quad (12)$$

where *P* is the incorrect detection probability performing energy detection, P_{toler} is the tolerated misidentify probability.

C. Weighted Bayesian Detection

When DC has received sensing reports, it needs to employ an appropriate fusion technique to make an accurate spectrum sensing decision. We apply a likelihood ratio test named WBD on data fusion. WBD is based on Bayesian detection [12], which is a hypothesis test for sequential analysis.

It requires the knowledge of prior probabilities of r_i 's when *r* is 0 or 1, i.e., $P[r_i|H_0]$ and $P[r_i|H_1]$. It also requires the knowledge of a prior probabilities of *r* , i.e., $P_0 = P[r=0]$ and $P_1 = P[r=1]$, which can be acquired with the method proposed in [7].

WBD can be represented by the following test, which inputs the sensing reports r_i of neighbors i and outputs a final spectrum sensing decision Γ.

$$\Gamma = \prod_{i=0}^{m} \left(\frac{P[r_i \mid H_1]}{P[r_i \mid H_0]} \right)^{V_i}$$
$$\begin{cases} \Gamma \geq \lambda \Rightarrow & accept\ H_1 \\ \Gamma < \lambda \Rightarrow & accept\ H_0 \end{cases} \quad (13)$$

where λ is a threshold calculated from

$$\lambda = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \quad (14)$$

where $C_{jk}(j = 0,1; k = 0,1)$ is the cost of declaring H_j true when H_k is present.

IV. SIMULATION

A. Simulation Environments

The simulation was run in MATLAB, and the same system environment with [7] was deployed to obtain comparable simulation results. The only difference was that we realized the simulation with multi-channel model. We compared three kinds of data fusion schemes, i.e., Bayesian detection, WBD and WSPRT [7]. We consider an ad hoc CR network with one PU as well as *N* SUs, among which *M* SUs are malicious. The primary user, a TV tower has twenty 6MHz channels in TV band, and the duty cycle of all the channels is fixed at 0.2. It locates D meters away from the center of the CR network. *N* SUs locate in a 2000m×2000m square area randomly, and follow a random waypoint movement model with a maximum speed of 10m/s and a maximum idle time of 120s. The transmission range of SUs is 250m. Three types of malicious nodes (always-busy, always-false and always-free) are same with normal SUs except reporting forged sensing reports. The layout of the simulated network is shown in Figure 1.

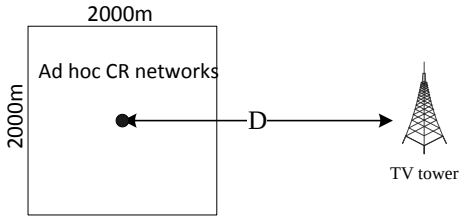


Figure 1. Simulation layout.

We use the HATA model for rural environments to calculate the path loss [11]. The values of the system parameters are listed in Table 1.

TABLE I. VALUES OF PARAMETERS USED IN THE SIMULATION

Parameter	Value
D	3000m
N	300
M	10,20,…,100
β	0.25
V_T	0.4
P_{toler}	0.01
P	0.3
PU antenna height	100m
SU antenna height	1m
transmitter power	100kW
receiver sensitivity	-94dbm
noise power	-106dbm

B. Simulation Results and Analysis

The threshold λ of Bayesian detection and WBD is calculated from (14), we first assume the perfect knowledge of P_0 and P_1, i.e., $P_0 = 0.8$ and $P_1 = 0.2$. The costs are assigned as: $C_{00} = C_{11} = 0$, $C_{10} = 1$, and $C_{01} = 10$. With these values, we can get $\lambda = 0.4$. Because the accurate knowledge on P_0 or P_1 may not be available, we simulated other threshold $\lambda' = 4\lambda = 1.6$. Another simulated fusion technique is WSPRT, the values of the parameters are the same with [7] except we deploy larger proportion of MUs.

We compare the performance of the three data fusion techniques. The metrics are correct sensing ratio, miss detection ratio and false alarm ratio, which add up to one. So we just focus on the first two metrics.

The number of malicious nodes increased from 0 to 100 at an interval of 10 in the three different attacks. Figures 2-4 show the simulation results when we consider always-busy, always-false and always-free attacks respectively. In all case, the correct sensing ratios of three types of data fusion techniques are more than 90% when the number of attackers is less than 30, which is acceptable based on the regulation of IEEE 802.22. But the performances diverge severely for the Bayesian detection with the number of MUs increasing, while the WBD is the most robust against attacks. The correct sensing ratios are above 80% with our proposed WBD under three types of attacks even the proportion of MUs is close to 1/3, while the miss detect ratios are acceptable at the same time. This shows that the trust model-based weight scheme has taken effect. WSPRT [7] employs similar weighted scheme with different trust evaluation strategy. It can reduce false information to a certain extent. But the increasing of malicious proportion would disturb the weight assignation in WSPRT, and finally, it largely increases the false alarm ratio.

It can be observed in Figure 4(a) that all the data fusion techniques perform stable under always-free attack, which increase miss detection ratio and decrease false alarm ratio. Figure 4(b) shows that the miss detection ratio is larger than other attacks.

 ISBN:978-1-61208-113-7

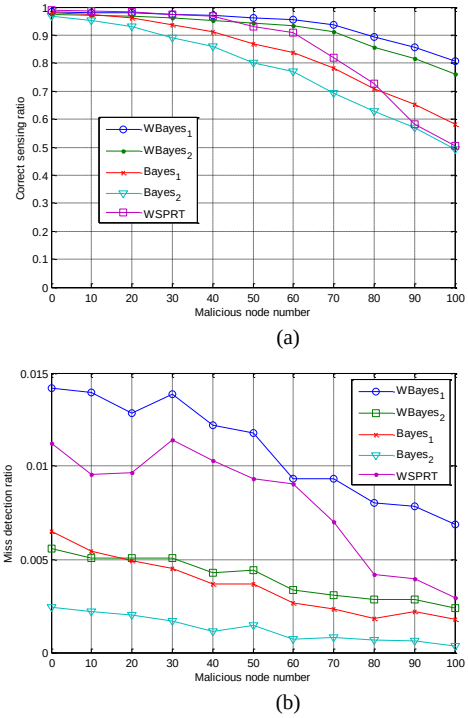


Figure 2. The performance of three fusion techniques with different number of always-busy attackers: (a) correct sensing ratio, (b) miss detection ratio. $\lambda = 0.4$: WBayes1, Bayes1; $\lambda' = 1.6$: WBayes2, Bayes2

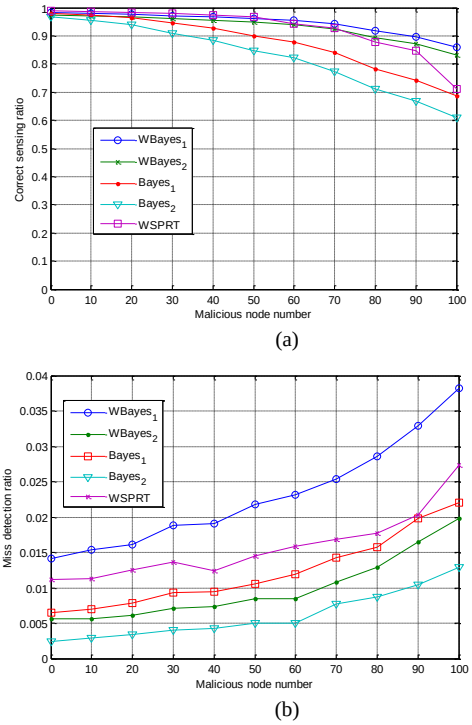


Figure 3. The performance of three fusion techniques with different number of always-false attackers: (a) correct sensing ratio, (b) miss detection ratio.

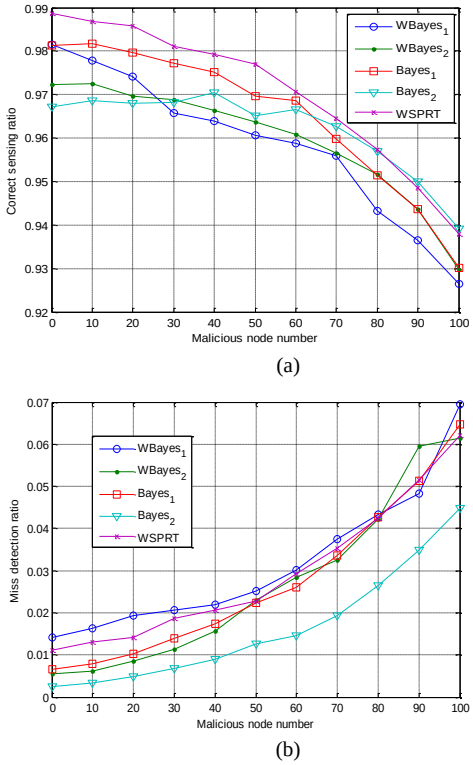


Figure 4. The performance of three fusion techniques with different number of always-free attackers: (a) correct sensing ratio, (b) miss detection ratio.

V. CONCLUSION

In this paper, we design two analysis models to identify anomaly behaviors in cooperative sensing process, and a PID trust model is employed to assign the credit value of SUs which can trace and nullify the malicious nodes rapidly. Simulation results demonstrate that comparing with the existing method the proposed scheme performs better especially in the case that there exist a large number of malicious nodes. In the behavior analysis model and data fusion technique the prior probability values play a key role, but the calculation of that needs many priori messages about the CR networks which may limit the deployment of the secure cooperative sensing techniques.

ACKNOWLEDGMENT

This work is supported by China NSF project No.61070211.

REFERENCES

[1] J. Mitola, "*Software Radio Architecture,*" John Wiley & Sons, 2000.

[2] Ian F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran, and Shantidev Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks : A survey," Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 50, Issue 13, September 2006, pp. 2127-2159.

[3] J. Unnikrishnan and V. Veeravalli, "Cooperative spectrum sensing and detection for cognitive radio, " in *IEEE Global*

Telecommunications Conference, GLOBECOM, Nov. 2007, pp. 2972 −2976.

[4] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunistic access in fading environments," in *Proc. IEEE Symp. New Frontiers in Dynamic Spectrum Access Networks (DySPAN'05)*, Baltimore, USA, Nov. 2005, pp. 131−136.

[5] S. M. Mishra, A. Sahai, and R. Brodersen, "Cooperative sensing among cognitive radios," in *Proc. IEEE Int. Conf. Commun.*, Turkey, June 2006, vol. 4, pp. 1658−1663.

[6] G. Ganesan and Y. G. Li, "Cooperative spectrum sensing in cognitive radio−part I: two user networks," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 2204−2213, June 2007.

[7] R. Chen, J. M. Park, and K. Bian, "Robust distributed spectrum sensing in cognitive radio networks," *Proceedings, INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, Apr. 2008. 1876-1884.

[8] P. Kaligineedi, M. Khabbazian, and V. K. Bharava, "Secure Cooperative Sensing Techniques for Cognitive Radio Systems," IEEE International Conference on Communications, May 2008, pp.3406-3410.

[9] T. Zhao and Y. Zhao, "A New Cooperative Detection Technique with Malicious User Suppression", IEEE International Conference on Communications, June 2009, pp,1–5.

[10] Z. Zhang, W. Jiang, and Y. Xue, "A Trust Model Based Cooperation Enforcement Mechanism in Mesh Networks", Proc. of the 6th International Conference on Networking (ICN), 2007, pp, 28-33.

[11] T. S. Rappaport, Wireless communications: principles and practice, vol.201. Prentice Hall PRT New Jersey, 1996.

[12] L. Lu, S.-Y. Chang, J. Zhang, L. Qian, J. Wen, V. K. N. Lau, R. S. Cheng, R. D. Murch, W. H. Mow, and K. B. Letaief, Technology Proposal Clarifications for IEEE 802.22 WRAN Systems, Mar. 2006. available at: http://www.ieee802.org/22/. [retrieved: September 14, 2010].

 ISBN:978-1-61208-113-7

Customer Security Concerns in Cloud Computing

Shirlei A. de Chaves, Carlos B. Westphall, Carla M. Westphall, and Guilherme A. Gerônimo
Network and Management Laboratory
Federal University of Santa Catarina
Florianópolis – SC - Brazil
Emails: {shirlei, westphal, carla, arthur}@lrg.ufsc.br

***Abstract*—There is no consensus about what exactly cloud computing is, but some characteristics are clearly repeated. It is a new distributed computing and business paradigm. It provides computing power, software and storage and even a distributed data center infrastructure on demand. In this paper, we investigated what are the main security concerns faced by the customers that are trying to better understand or profit from this new paradigm, especially considering a public cloud and we conclude that data confidentiality, integrity and availability are the biggest ones.**

Keywords - Cloud computing; security; distributed computing.

I. INTRODUCTION

Despite of the fact that industry big players like Google, Amazon, SalesForce, Microsoft and others have products and services under the umbrella of 'cloud computing', 'cloud ready' or other similar denomination, there is no consensus about what exactly cloud computing is. Below we list some definitions made by researchers:

"*Cloud computing is the next natural step in the evolution of on-demand information technology services and products. To a large extent cloud computing will be based on virtualized resources.(…) Cloud computing embraces cyber infrastructure and builds upon decades of research in virtualization, distributed computing, grid computing, utility computing, and more recently networking, web and software services* [1]."

"*A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet* [2]."

"(…) *cloud computing is a nascent business and technology concept with different meanings for different people. For application and IT users, it's IT as a service (ITaaS) - that is, delivery of computing, storage, and applications over the Internet from centralized data centers. For Internet application developers, it's an Internet-scale software development platform and runtime environment. For infrastructure providers and administrators, it's the massive, distributed data center infrastructure connected by IP networks* [3]."

""*A Cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers* [4]."

"*a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction* [5]."

From these definitions, it is possible to notice that some characteristics are clearly repeated. It is a new paradigm, not just a distributed computing paradigm, but also a new business paradigm. It is intended to provide computing power, software and storage and even a distributed data center infrastructure on demand. In order to make these characteristics viable, cloud computing makes use of existing technologies, such as virtualization, distributed computing, grid computing, utility computing and Internet. However, even those industry big players have products and services available as also a definition of what are the basic cloud computing underlying technologies, a customer intending to better understand and profit from this new paradigm faces several concerns, especially the ones related to security.

Considering the customer point of view, we have made an extensive research to obtain what are the main security problems pointed in the available literature for cloud computing security, aiming to list and discuss the more recurrent ones. The results and the discussion are presented in Section 3. It is also worth to mention that being the cloud computing security subject under active research, many changes, new events or studies relating to it are coming out in a rapid pace, so this paper does not aim to exhaust it, but to contribute to the discussion.

.

II. CLOUD COMPUTING CATEGORIES

Attempts to cloud computing standardization are being done by some groups, including governments and industry. One effort that can help to avoid

 ISBN:978-1-61208-113-7

misunderstandings, by putting everyone to talk the same language, is the definition of cloud computing and its categories. As of this writing, the US National Institute of Standards and Technology (NIST) is one of them, having defined the cloud as composed of four deployment models, three service models and five essential characteristics. The Cloud Security Alliance [6], which formal debut was made at RSA Conference 2009 releasing a white paper entitled "Security Guidance for Critical Areas of Focus in Cloud Computing", has taken these definitions to work through its guidance, explaining that the motivation is "to bring coherence and consensus around a common language so we can focus on use cases rather than semantic nuance [6]".

A. *Deployment models*

The definitions of the deployment models listed next are taken as it is from the NIST definition, although other researches mention this deployment models with similar definitions.

1) *Public Cloud*

The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

2) *Private Cloud*

The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party.

3) *Hybrid Cloud*

The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds).

4) *Community Cloud*

The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party.

B. *Service Models*

The three service models listed below are not exclusively from NIST, being mentioned in several papers, including [2][5].

1) *Infrastructure as a Service (IaaS)*

"The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications [5]."

2) *Platform as a Service (PaaS)*

"The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider [5]."

3) *Software as a Service (SaaS)*

In this case, is provided "a complete, turnkey application—including complex programs such as those for CRM or enterprise-resource management via the Internet [5]." Or, in the words of NIST, "the capability provided to the consumer is to use the provider's applications running on a cloud infrastructure [5]."

In each of these service models, what can be controlled by the customer varies, but in general, he does not have control over the underlying cloud infrastructure. This is especially true when is the case of a public cloud, the focus of the present paper. In a private cloud, for example, security responsibilities can be taken on by the customer, if he is managing the cloud, but in the case of a public cloud, such responsibilities are more on the cloud provider and the customer can just try to assess if the cloud provider is able to provide security.

The five essential characteristics are related to characteristics already mentioned in the introduction of this paper: on-demand self-service, broad network access, resource pooling, rapid elasticity and measured services.

III. CLOUD COMPUTING SECURITY OVERVIEW

Many cloud computing security problems are still unclear. Being cloud computing such a recent computing paradigm, it is natural that many aspects remain uncovered whereas the paradigm itself is being more developed and understood.

According to [5], there are three main customers' concerns:

Vulnerability to attack: critical business information and IT resources are outside the customers firewall.

Standard security practices: customers want to be confident that such practices are being followed. Most of those practices require disclosure and inspection, which leads to another concern as a customer: will my data be in the same virtual hardware and network resources with other customers, being susceptible to disclosure in someone else's inspection?

Being subject to state or national data-storage laws related to privacy or record keeping: European Union (EU), for example, has privacy regulations that do not

 ISBN:978-1-61208-113-7

permit some personal data to be transmitted outside the EU. In the cloud, data can be stored anywhere in the world; it is important to attend such regulations.

In June 2008, the Gartner Group released a report entitled "Assessing the Security Risks of Cloud Computing" [7]. According to this report, widely commented and cited on the Internet, before jumping into the cloud, the customer should know its unique security risks, considering specially seven security conditions during the process of choosing a cloud provider. These unique security risks are:

Privileged user access: outsourcing means allowing outsourced services to bypass internal controls, including personnel controls. With this in mind, the customer has to obtain as more information as possible about how the possible future provider hires people and what kind of controls their accesses have.

Regulatory Compliance: if the cloud computing provider is not subject of external audits and security certifications, the customer probably should not use its services for non trivial tasks. Customers have to always remember that, unless stated or agreed otherwise, they are responsible for their own data.

Data location: when using the cloud, the customer probably will not know where their data will be stored. Thus, it is recommended checking if the provider will commit to store and process data in specific jurisdictions and if a contractual commitment on behalf of the customer will be made by the provider.

Data segregation: customers should check what is done to separate different customers' provider data, due to the fact that, in a cloud, the environment is shared. Using cryptography, for example, is effective, but do not solve all the problems. It must be checked also if the cryptographic schemes are designed and tested by specialists, because cryptographic accidents are able to make data unusable.

Recovery: the provider capacity of restoring the entire system and how long it would take should be checked by the customer. Any provider that does not replicate its data or infrastructure is prone to total failures.

Investigative support: In order to have confidence that inappropriate or illegal activities will be possible to be investigated, the customer needs a formal commitment from the provider. This commitment should state which kind of investigation will be possible and also gives evidence that similar support was already done by the provider. Otherwise, the customer almost can be sure that such investigations will be impossible.

Long-term viability: if happens that the cloud computing provider be acquired or goes broke, the customer needs to know if the data will still be available and in a format that will allow being imported to a substitute application.

Summarizing the Gartner's report, customers should demand transparency and avoid providers that do not offer clear information about security programs.

In a Tech News from forbes.com, published online in February 02, 2008 [8], by Andy Greenberg is cited that when customers store their data in someone else's software and hardware, "they lose a degree of control over their often-sensitive information". In that article, Greenberg gives the example of an employee of an investment bank that uses Google Spreadsheets to organize a list of bank employees and their social security number. In this example, the responsibility of protecting such information from hackers and internal data breaching is not from the bank, but Google's. Another situation in this same case is that, if government investigators subpoena Google to supply that list, sometimes even with no customer knowledge, Google may attend. Google's privacy politics says that it will share data with the government if it has a "good faith belief" that this is necessary [9]. Greenberg also points that other problems to cloud computing is the cyber crimes. He gives some examples occurred in 2007, like:

- "retailer TJX lost 45 million credit card numbers to hackers";
- "the British government misplaced 25 million taxpayer records";
- "software company Salesforce.com sent a letter to a million subscribers describing how some customers' e-mail addresses and phone numbers had been snagged by cybercriminals - and warning how another wave of phishers were attempting to send malware more broadly to Salesforce.com customers".

Analyzing the articles cited before, it is possible to visualize that the main concerns, if not all, are related to the business level, i.e., customers are worried about how their business process will be affected. This situation is expected because cloud computing can also be seen as a new business model, with many aspects to be fully understood before being adopted with no restrictions or, better, with fewer restrictions. As happen in any new business or technological area, customers and professionals need to be confident on what are they getting into. Foster [2] also considers this business model characteristic in cloud computing, saying that "as for Utility Computing, it is not a new paradigm of computing infrastructure; rather, it is a business model in which computing resources, such as computation and storage, are packaged as metered services similar to a physical public utility, such as electricity and public switched telephone network." Also, it is possible to visualize that major concerns are about data: how and where it will be kept, who will be able to access it and which regulations

 ISBN:978-1-61208-113-7

will have it as subject. Considering that, data confidentiality, integrity and availability will be discussed in more details in the next section.

A. Confidentiality, Integrity and Availability (CIA): the big concern

According to [11],"storing data remotely into the cloud in a flexible on-demand manner brings appealing benefits: relief of the burden for storage management, universal data access with independent geographical locations, and avoidance of capital expenditure on hardware, software, and personnel maintenances, etc." However, users face the situation of losing control of their data. For example, he no longer has physical possession of the outsourced data and may not get to know about data loss and leakage incidents, if the cloud provider for some reason acts unfaithfully and decide not to report the incident [11], just to cite a few. Having that in mind, in the remainder of this section we discuss some of the traditional ways of delivering data confidentiality, integrity and availability.

Cryptography

One could ask if applying some cryptographic and backup schema would not solve at least part of the problem. This is a question certainly being target of studies, especially because as we cited before, cryptographic accidents are able to make data unusable. Trying to contribute to the subject, we bring some questions to this discussion.

- If using cryptography, how the key management is done?
  - One key for each customer?
  - One key to all customers?
  - Multiple keys for the same customer?
- What are the current cryptographic systems more applicable to the cloud computing characteristics, especially data storage?
- Last but not least, in which situations cryptography should be used?

We think that the cloud provider should have a detailed cryptographic plan, explaining what algorithms will be used, how the key management will be done, when encryption will be used and so on. The Cloud Security Alliance Guide [6] provides some guidance in these questions. As stated by them, cloud computing divorces components from location and this creates security issues that result from this lack of any perimeter. Hence there is only one way to secure the computing resources: strong encryption and scalable key management. Also according to [6], cloud customers and providers must encrypt all data in transit, at rest or on backup media, since all communications and all storage may be visible to arbitrary outsiders. Customers and providers want to encrypt their data to ensure integrity and confidentiality as also to avoid having to report incidents to their users (remembering that a provider's customer may have their own customer to report and successively).

According to [10], users are "universally required to accept the underlying premise of trust.", highlighting that although some take trust as synonymous of security, it is not and in security the element of trust is more apparent. Relating to the classic key concepts of information security, the CIA, [10] lists the minimum capabilities that should be offered by the cloud storage provider:

• "a tested encryption schema to ensure that the shared storage environment safeguards all data;

• stringent access controls to prevent unauthorized access to the data; and

• scheduled data backup and safe storage of the backup media [10]."

Wang [11] proposes public auditability for cloud data storage security. Such audit would be done by a third party auditor, called TPA. Knowing that such data in general, due to privacy issues, cannot be subject of disclosure, [11] lists two fundamental requirements for the TPA: 1) efficient cloud data storage auditing without demanding the local copy of data and without additional on-line burden to the cloud user; 2) no new vulnerabilities should be brought to user data privacy by the auditing process. Such requirements are best practices as also are the reasons [11] mention for the TPA being a good choice instead of the own user auditing the correctness of their data: 1) possible large size of stored data; 2) possible user' computer resource constraints; 3) "simply downloading the data for its integrity verification is not a practical solution due to the expensiveness in I/O cost and transmitting the file across the network."

Backup and recovery

Backup is probably the more traditional way of keeping data for recovery purposes. However, being crucial to ensure that a point-in-time data is available to restore business operations and given the special nature of a cloud environment, some questions need to be clearly answered by the provider and understood by the customer:

- Who performs the backup?
- How frequent the backup is performed?
- Who is responsible for storing the backup?
- Which backup format is used? Is it dependent of a specific technology?
- Logical segregation of data is maintained through the backup execution?

Having these questions being done, another important issue is if the provider will be able to meet any specific customer backup requirement. Normally, to have an

 ISBN:978-1-61208-113-7

effective backup and recovery strategy, a careful study of organization's need have to be done. Being the cloud a multi-tenant environment, it is possible that the cloud provider specific backup and recovery plan will not fit completely to the customer's need. Also, as mentioned before, the data should be encrypted on the backup media. According to [6], as a customer and provider of data, it is customer's responsibility to verify that such encryption takes place.

B. Data format standards

It seems vital to data availability to have a data format that allows customers to take their data from one provider and leverage it inside another provider's application. This kind of concern, however, is neither new or exclusively of cloud computing, so what was already learnt or developed since the beginning of the Internet and the need of data exchange should be taken into account when addressing this situation. Some standardizations initiatives are in progress, like the Cloud Computing Effort announced on April 27, 2009 by DMTF (Distributed Management Task Force) [12].

We do not know what would be the better in terms of data storage specifically and this is not the focus of this discussion. Maybe dictating a specific format is not a viable idea, at least not in a short time. But, the data interchange should be specified in some standard or well accepted format. The XML (Extensible Markup Language) format was designed to store and transport data. As cited in [13], XML is a technology that started a decade ago and since then great effort has been done by the research and industrial community to support XML and related technologies in RDBMS (Relational Database Management System). Also, being the format subject of standardization and widely adoption, lots of research aiming to secure the format has already been developed and it still is subject of ongoing improvements.

Adopting XML or not, the groups cited here and many others working on cloud computing standardizations should have this in mind: data must be interchangeable.

IV. FINAL CONSIDERATIONS

Maybe the cloud will evolve and become the largest information system we ever saw, having all sort of data and dealing with all kind of information, all kind of sensitive information. So, much research work is in progress to provide security for cloud computing, especially regarding do data confidentiality, integrity and availability. The general believe, including ours, is that the larger adoption of cloud computing relies on how secure it is and that security should be addressed since the very beginning. Being cloud computing a still evolving paradigm, some new security concerns may appear during the definition process, but the concerns highlighted in the present survey probably will not change. There are, however, a lot of good research and work in progress aiming to mitigate or to solve the security issues and to turn the cloud computing horizon less cloudy. Among these researches are government initiatives, like the cloud security group from US National Institute of Standards and Technology (NIST) and industry initiatives, like the Cloud Computing Security Alliance. Having data confidentiality, integrity and availability a strong legal side, some legal organizations like Strafford Publications are organizing events to discuss the subject, like a Teleconference entitled "Cloud Computing: Managing the Legal Risks" [14], showing that other areas beyond information technology are watching cloud computing growing adoption more closely. Such initiatives bring advantages for the customer that can have more qualified background when analyzing the available cloud computing solutions to migrate his services to a cloud.

REFERENCES

[1] M. A. Vouk, "Cloud Computing – Issues, research and implementations", In: 30th International Conference on Information Technology Interfaces, pp. 31-40, 2008.

[2] I. Foster, I. Yong Zhao Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-Degree Compared", In: 2008 Grid Computing Environments Workshop, pp. 1-10, 2008.

[3] G. Lin, D. Fu, J. Zhu, and G. Dasmalchi, "Cloud Computing: IT as a Service," IT Professional, vol. 11, no. 2, pp. 10-13, Mar./Apr. 2009.

[4] R. Buyya, Y. Chee Shin, S. Venugopal, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities", In: 10th IEEE Conference on High Performance Computing and Communications. IEEE Computer Society, 2008, pp. 5-13.

[5] P. Mell, and T. Grance, "Cloud computing definition". NIST, June 2009, http://csrc.nist.gov/groups/SNS/cloud-computing/index.html

[6] Security Guidance for Critical Areas of Focus in Cloud Computing, www.cloudsecurityalliance.org (last access on Dec. 2010).

[7] J. Heiser, and M. Nicolett, Assessing the Security Risks of Cloud Computing. http://www.gartner.com/DisplayDocument?id=685308

[8] http://www.forbes.com/technology/ (last access on Dec. 2010).

[9] http://www.google.com/privacy/privacy-policy.html (last access on Dec. 2010).

[10] L. M. Kaufman, "Data security in the world of cloud computing," IEEE Security & Privacy Magazine, vol. 7, no. 4, pp. 61-64, July 2009. [Online]. Available: http://dx.doi.org/10.1109/MSP.2009.87

[11] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," in *2010 Proceedings IEEE INFOCOM*. IEEE, March 2010, pp. 1-9.

[12] http://www.dmtf.org/ (last access on Dec. 2010).

[13] R. Zhen Hua Liu Murthy, "A Decade of XML Data Management: An Industrial Experience Report from Oracle", In: IEEE 25th International Conference on Data Engineering, pp. 1351-1362. IEEE Computer Society, 2009.

[14] http://www.straffordpub.com/products/cloud-computing-managing-the-legal-risks-2009-12-09 (last access on Dec. 2010).

 ISBN:978-1-61208-113-7

Access Control in a Form of Active Queuing Management in Congested Network Environment

Vladimir Zaborovsky
St. Petersburg state Polytechnical University
Saint-Petersburg, Russia
e-mail: vlad@neva.ru

Vladimir Mulukha
St. Petersburg state Polytechnical University
Saint-Petersburg, Russia
e-mail: vladimir@mail.neva.ru

***Abstract* — Internet processes information in the form of distributed digital resources, which have to be available for authorized use and protected against unauthorized access. The implementation of these requirements is not a simple task because there are many ways to its realization in the modern multiserviced and congested networks. In this case many well-known solutions of the past became inappropriate because of traffic fractal statistics, which are caused by persistent packet dynamics of transport protocols and loss of available throughput. Therefore we offer the new approach to raise access control functionality, taking into account models of transport protocols in congested network environment, characteristics of virtual channel throughput and features of active queuing management mechanism that based on randomized preemptive procedure.**

Keywords* — *access control, authorized use, virtual connection, priority queueing management, randomized push-out mechanism

I. INTRODUCTION

Internet as a global information infrastructure is used widely for business, education and research. This infrastructure keeps information in the form of distributed digital resources that have to be available for authorized use, and protected against unauthorized access. However, the implication of these requirements is not a simple task due to many elements and many ways of realization. Therefore solutions of the past have become inappropriate because of traffic fractal statistics, which are caused by persistent packet dynamics and correspondent loss of virtual channel available throughput. In this paper we propose a new approach to access control flexibility enhancement based on active queuing management mechanism and randomized preemptive procedure. The offered solution can be implemented by a firewall and can be applied in the existing network environments.

To reach this purpose we propose: 1) the new classification of virtual connections (VC) based on security characteristics and throughput requirements; 2) VC model, which takes into account fractal characteristics of packet flows; 3) randomized preemptive queuing management mechanism in congested networks. We use a combined method of VCs throughput management that unites principles of feedback and program control within a framework for Policy-based Admission Control (Fig. 1):

- Policy Decision Point (PDP).
- Policy Enforcement Point (PEP) – security-critical component, which protects the resources and enforces the PDP's decision.
- Policy Administration Point (PAP).

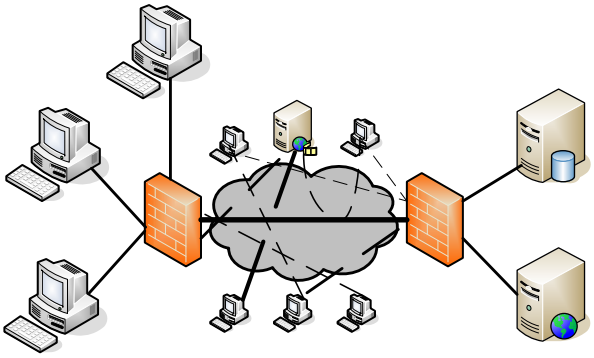


Figure 1. Firewall as a central component of access policy enforcement

In this framework firewall combines PDP and PEP by controlling access request and enforcing access decisions in real-time. In this case, access control can be considered as the throughput control of VC. So, access to the specific network resource is prohibited if the corresponding VC between the user and resource has no available throughput. Therefore from PAP firewall receives two types of access policy rules: packet filtering rules and data flow rules.

The parameters of firewall rules depend on the set of network environment and/or protocols characteristics *A*. This set can be divided on two classes with different access conditions. In proposed approach the classification decision is based on indicator function *F* and firewall has two modes in accordance to possible *F (A)* values:

- 0, if the data flow is forbidden according to the access policy (filtering rules);
- 1, if the data flow is permitted.

Forbidden mode means that access denied by PAP. Then the subset of permitted flows is divided into new two subsets:

- priority ones that have low throughput and demand low stable delivery time;
- background ones that demand high throughput and has no delivery time requirements.

To provide this classification procedure we proposed active queuing management mechanism, which based on randomized preemptive control. Therefore in the firewall the data flow throughput and time that packets spend in queue (minimum value for priority permitted flows and infinity for denied) are the functions of randomized control parameter α. Each of the firewall rules has a set of attributes: identifiers of subject and object and the access

 ISBN:978-1-61208-113-7

rights from one to another. In the modern network environment access rules have much more attributes that need to identify two subsets of permitted flows. Therefore the actual problem of access control within framework for Policy-based Admission Control is the flexible configuration of firewall rules, which considers dynamics of network environment including specific congested conditions. In this paper we introduce active queuing management mechanism for access control policy enforcement based on randomized preemptive procedure and network environment characteristics.

The paper is organized as follows: In Section II we suggest new classification of virtual connections. In Section III the model of virtual connection is presented. The Section IV and V are the theoretical parts of the paper where the mathematical model and basic equations are analyzed and estimated. The Section VI is about practical usage of proposed method.

II. VIRTUAL CONNECTION CLASSIFICATION

In this paper we use the term "access management" as the combine of access control and traffic management. Access control is the basic technical method of information security in the computer networks. It is providing confidentiality by blocking the denied data flows, availability by permitting legal connections and integrity by reducing the risk of data modification or destruction. Confidentiality, integrity and availability are the core principles of information security. Access control is based on subject-object model, where subjects are the entities that can perform actions in the system and influence the environment condition and objects are the entities representing passive elements between which access need to be controlled. Data flows between objects and subjects named virtual connections (VC). In this paper Virtual Connection is the type of information interaction between applications on object and subject by means of formation one-way or duplex packet stream, and also the logical organization of the network resources necessary for such interaction.

Computer network can be considered as the set of such VC. In classical subject-object model the set of VC is divided into two subsets:

- Non forbidden connections that do not harm the protected information;
- Forbidden connections that can low the confidentiality, integrity or availability of protected information.

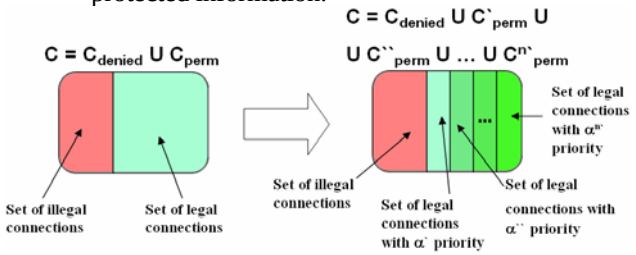


Figure 2. Virtual connections classification model.

We consider spreading the set of legal connections into several subsets by priority characteristics. In this paper we present the simplest example with two subsets:

- Non forbidden priority connections;
- Non forbidden non priority or background connections.

On Fig. 2 there is graphical interpretation of considered classification.

III. VIRTUAL CONNECTION MODEL

The modeling of the VC behavior has received considerable attention in recent years. In this paper we present a simple model of VC. Each connection can be described by several parameters:

$$Vc(S,O,Th,Type,Fr)$$

where S,O are the subject and object of information interaction, Th – virtual connection throughput, $Type$ – the resource requirements, Fr – fractal nature of VC.

From this point of view we suggest to divide set of virtual connections into two subsets:

- Fractal natured virtual connections based on transport protocols with feedback (TCP connections)
- Data flows without fractal properties like UDP data streams

Researches have shown that fractal properties of VC influence its throughput. For calculation the average throughput of TCP connection it is necessary to create a model of connection with fractal properties.

In this paper we suggest to use a simple discrete time model of TCP connection: at each discreet time moments "*k*" TCP throughput "*Th*" can be describes by formulas:

$$X_{k+1} = R(A, X_k, \xi_k)X_k, \quad Th_k = F(X_k),$$

where X – congestion window, which size measures in conventional unit, A – vector of the protocol deterministic characteristics; ξ - stochastic variable describes by density distribution function [3][4]

$$R(A, X_k, \xi_k) = \begin{cases} 1; \xi_k = 0, X_k = C \\ 1/2; \xi_k = 1 \\ 1/X_k; \xi_k = 2 \\ 2; \xi_k = 0, X_k < C, X_k < S \\ (X_k + 1)/X_k; \xi_k = 0, X_k < C, X_k > S \end{cases}$$

where C is TCP receive window size, S – threshhold.

As it is known from an example of Cantor set the fractal properties appears at loss of the set's part. Fractal properties of TCP-connection characterize the throughput losses because of feedback mechanism. On Fig. 3 there are shown the throughput losses because of CWND adaptation mechanism.

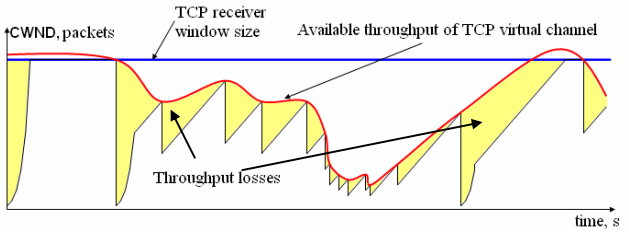


Figure 3. TCP throughput losses because of CWND mechanism.

We suggest using different algorithms to calculate the throughput of VC with fractal properties and without ones.

For the connections without fractal properties we will use the simple formula:

$$Th = Th_0 \cdot (1 - p),$$

where Th_0 is the connection throughput from the stream source and p is the packet loss probability.

For TCP connections we use the well-known formula:

$$Th = \min\left(\frac{C}{RTT}; \frac{1}{RTT \cdot \sqrt{\frac{2}{3}p}}\right),$$

where C is TCP receive window size, RTT is round trip time and p is the packet loss probability (loss rate). The graph of this function for C = 100 packets and RTT=110 ms is shown on Fig. 4.

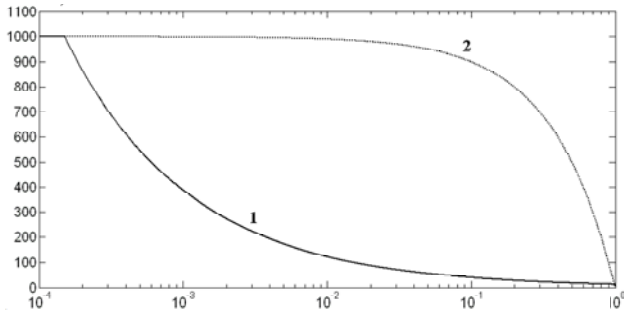


Figure 4. Dependence of TCP throughput on packet loss probability for TCP connections.

IV. MODEL OF NETWORK ENVIRINMENT

According to the VC models written above we consider the preemptive priority queueing system with two types of customers. First type of customers has priority over the second one. The customers of the type 1 (2) arrive into the buffer according to the Poisson process with rate λ_1 (λ_2). The service time has the exponential distribution with the same rate μ for each type. The service times are independent of the arrival processes. The buffer has a finite size k $(1 < k < \infty)$ and it is shared by both types of customers. The absolute priority in service is given to the customers of the first type. Unlike typical priority queueing considered system is supplied by the randomized push-out mechanism that helps precisely and accurate to manage customers of both types. If the buffer is full, a new coming customer of the first type can push out of the buffer a customer of type 2 with the probability α. We have to mention that if α =1 we retrieve the standard non-randomized push-out.

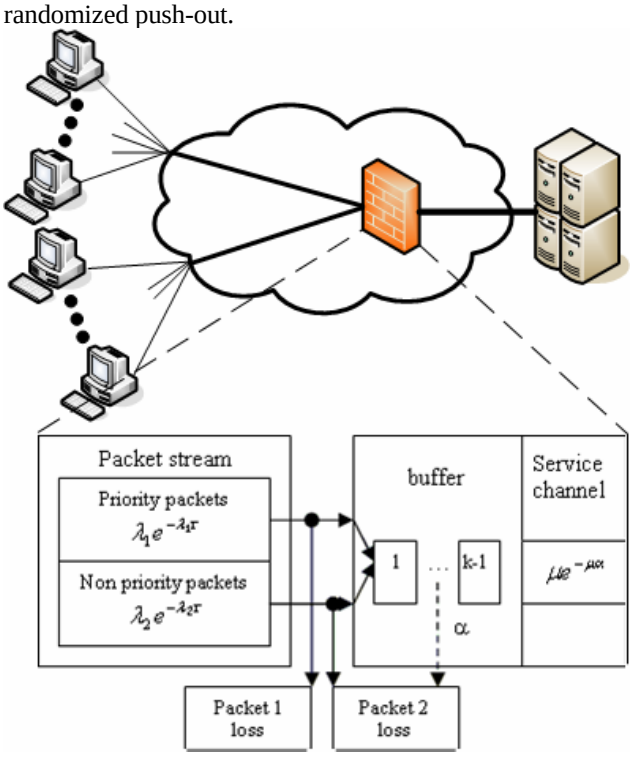


Figure 5. Priority queueing schema $\vec{M}_2 / M / 1 / k / f_2^1$ of telematics network devide.

The scheme described priority queueing is resulted on Fig. 5. The priority queueing without the push-out mechanism ($\alpha = 0$) and with the determined push-out mechanism ($\alpha = 1$) are well-studied. The concept of the randomized push-out mechanism with reference to network and telecommunication problems is offered in [1] where this mechanism was combined with relative priority, instead of absolute, as in our case.

The summarized entering stream represented on Fig. 5 will be the elementary with intensity $\lambda = \lambda_1 + \lambda_2$. The priority queueing represented on Fig. 5, is $\vec{M}_2 / M / 1 / k / f_2^1$ type by Kendall's notation.

Problems of research priority queueing have arisen in telecommunication with the analysis of real disciplines of scheduling in operating computers. Last years a similar sort of queueing model, and also their various generalisations are widely used at the theoretical analysis of Internet systems.

As shown in [1], the probability pushing out mechanism is more convenient and effective in comparison with other mathematical models of pushing out considered in the literature. It adequately describes real processes of the network traffic and is simple enough from the mathematical point of view. The randomized

 ISBN:978-1-61208-113-7

push-out mechanism helps precisely traffic management and security. The another control and security factor is the telematics device buffer size. It can be varied to increase the throughput of necessary connections and reduce throughput of suspicious ones.

V. MAIN EQUATIONS

The state graph of system $\vec{M}_2 / M / 1 / k / f_2^1$ is presented on Fig. 6.

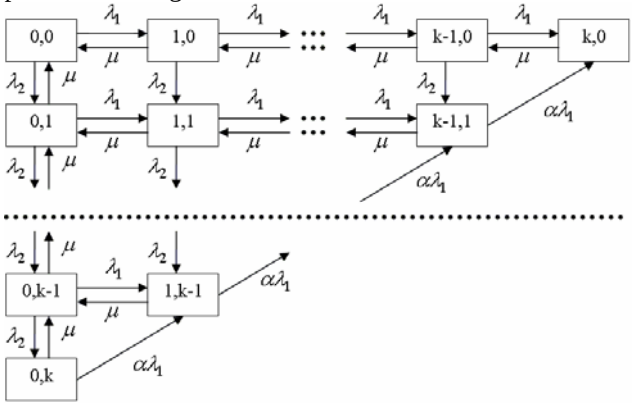


Figure 6. The state graph of $\vec{M}_2 / M / 1 / k / f_2^1$ type system.

Making by usual Kolmogorov's rules set of equations with the help of state graph we will receive:

$$-[\lambda_1(1-\delta_{j,k-i})+\alpha\lambda_1(1-\delta_{j,k})\delta_{j,k-i}+\lambda_2(1-\delta_{j,k-i})+$$
$$+\mu(1-\delta_{i,0}\delta_{j,0})]p_{ij}+\mu p_{i+1,j}+\mu\delta_{i,0}p_{i,j+1}+\lambda_2 p_{i,j-1}+ \quad (1)$$
$$+\lambda_1 p_{i-1,j}+\alpha\lambda_1\delta_{j,k-i}p_{i-1,j+1}=0,(i=\overline{0,k};\ j=\overline{0,k-i}),$$

where $\delta_{i,j}$ is the delta-symbol.

There is a normalization condition for the system:

$$\sum_{i=0}^{k}\sum_{j=0}^{k-i} p_{ij}=1.$$

At real k (big enough) this system is ill-conditioned, and its numerical solution leads to the big computing errors. In this paper we use the method of generating functions [1] in its classical variant offered by H.White, L.S.Christie and F.F.Stephan with reference to $\vec{M}_2 / M / 1 / f_2$ type systems.

Solving (1) system we receive some auxiliary variables [4]

$$p_i = p_{k-i,i},\ (i=\overline{0,k}),$$

$$q_{k-j}=(1-\alpha)\sum_{i=1}^{j} p_i\rho_1^{i-j}+q_k\rho_1^{-j},\ (j=\overline{1,k}),$$

$$r_n=\frac{(1-\rho)\rho^n}{(1-\rho^{k+1})},\ (n=\overline{0,k}).$$

When using them we can receive loss probability for priority ($P_{loss}^{(1)}$) and non-priority ($P_{loss}^{(2)}$) packets:

$$P_{loss}^{(1)}=q_k+(1-\alpha)\sum_{i=1}^{k-1} p_i,$$

$$P_{loss}^{(2)}=r_k+\alpha\frac{\rho_1}{\rho_2}\sum_{i=1}^{k} p_i.+\frac{\rho_1}{\rho_2}p_k$$

By these formulas we received some graphs for different rate of input streams of relative throughput of this type (Fig 7,8)

$$æ_i=1-P_{loss}^{(i)},\ (i=\overline{1,2}).$$

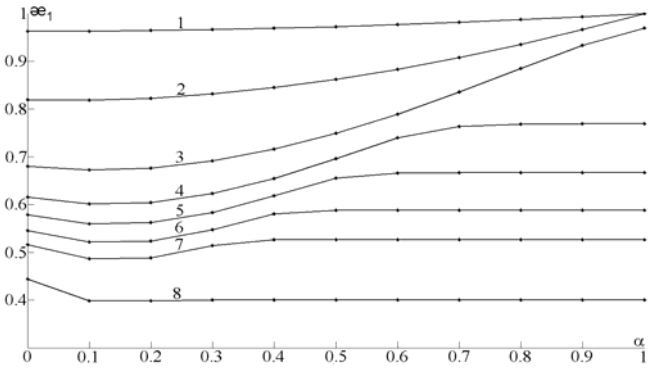


Figure 7. Relative throughput of priority packets for strongly congested transport virtual channel with $\rho_2=1,5$ and different values ρ_1: 1 – $\rho_1=0,1$; 2 – $\rho_1=0,5$; 3 – $\rho_1=1,0$; 4 – $\rho_1=1,3$; 5 – $\rho_1=1,5$; 6 – $\rho_1=1,7$; 7 – $\rho_1=1,9$; 8 – $\rho_1=2,5$. The same legend is used by all Figures.

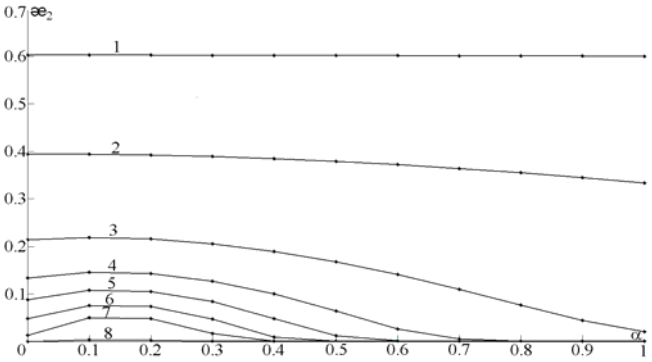


Figure 8. Relative throughput of non-priority packets

From Fig. 7 and 8 we can see, that by choosing parameter α, we can change $æ_i$ in very wide range. For some ρ_1 values variable $æ_i$ changes from 0.6 to 1 while $\lambda_1+\lambda_2 >> \mu$. There is an extremum on the most of the curves at $\alpha=0,1-0,2$. It means that increasing the push-out probability of non priority packets thus we reduce probability of their loss in the strongly congested networks. It can be explained by the fact that various mechanisms work in the absence of push-out mechanism ($\alpha=0$) and while $\alpha>0$.

The relative time that the priority packet spend in queueing can be calculated by Little's Formula (Fig 9,10):

$$\theta_i=\frac{\bar{s}_i}{\bar{\tau}_i}=\frac{\bar{n}_{load}^{(i)}}{(1-\bar{P}_{loss}^{(i)})}+\rho_i,\ \bar{\tau}_i=\frac{1}{\lambda_i},\ (i=\overline{1,2}).$$

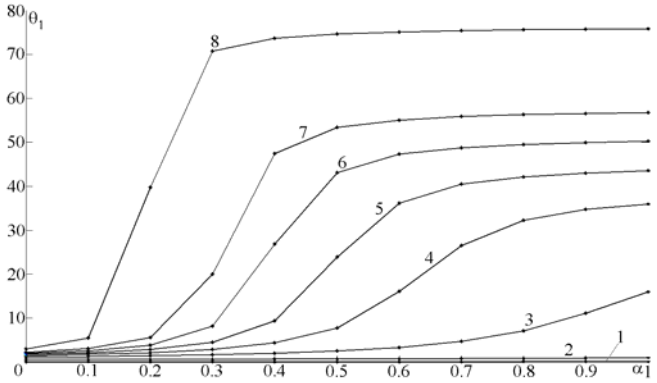


Figure 9. The time that priority packet spend in queueing

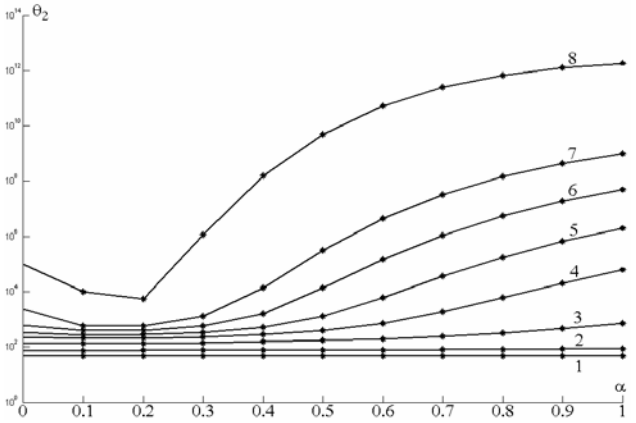


Figure 10. The time that non-priority packet spend in queueing

Fig 9,10 show that proposed queueing mechanism provide a wide range of control feature by randomized push-out parameter α and buffer size k. According to the packet's mark (Forbidden, Priority, Background) the period that packet spend in queue can vary from 1 to 10^{14} times, which can be used to control access to information resource providing confidentiality.

For highly congested network the priority type is much less important, than the push-out mechanism and the value of α parameter. The push-out mechanism allows to enforce access policy using traffic priority mechanism.

By choosing α parameter we can change the time that packets spend in the firewall buffer, which allows to limit access possibilities of background traffic and to block forbidden packets. So by decreasing the priority of background VCs and increasing the push-out probability α we can reduce the VC throughput to low level without interrupting it.

The most wide range of control can be reached in intermediate environment conditions when linear law of the losses has already been broken, but the saturation zone has not been reached yet. Numerical experiment [4] has been made to detect conditions in which ρ_1 varied over a wide range from 0,1 to 2,5, and $\rho_2 = 1{,}5$.

VI. PRACTICAL USAGE AND FUTURE DEVELOPMENT.

Good example of opportunity to use such mechanism is the problem of controlling removed robotic object, which telemetry data and a video stream are transmitted on global networks. In this case control commands are transmitted by TCP, and a video stream data are transmitted by UDP. A mean values of throughput of our robotic object: throughput of TCP channel (control and telemetry packets) ~100Kb/s, throughput of UDP video stream ~1,2Mb/s.

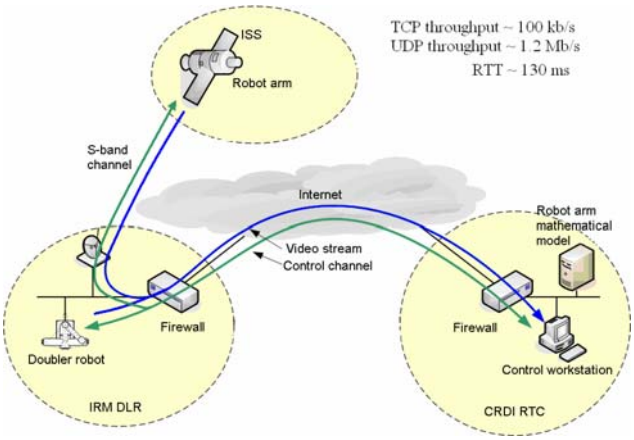


Figure 11. The scheme of space experiment "Contour"

In a considered example on Fig. 11 (ROKVISS mission [5]), the choice of a priority of service and loss-probability of a priority packet α allows to balance such indicators of functioning of a network, as loss-probability of control packets $p_{loss}^{(1)}$ and quality of video stream for various conditions of a network environment. The parameter α can vary for delay minimization in a control system's feedback.

The given problem is important for interactive control of remote real-time dynamic objects, in a case when the complex computer network is the component of a feedback control contour, therefore minimization of losses and feedback delays, is the important parameter characterizing an effectiveness of control system.

In future this method of preemptive access management could be used to mature the DTN technology for space exploration missions and communications architecture for example for robots control on planet's surface from orbital station through the network environment with unstable throughputs and unpredictable packet delays.

Of course in this case two types of priority are not enough for enforce access policy in multiservice network environment, but the recurrent mode of proposed procedure can increase the number of priority VC subsets.

 ISBN:978-1-61208-113-7

VII. CONCLUSION.

1. The offered access control approach allows more deeply and more detailed understanding of requirements of access policy in the form of firewall configuration rules.

2. Proposed model based on DiffServ approach considers computer network as the set of VCs, which throughput is easy controlled by proposed classification procedure and algorithm that divides the set of non forbidden VCs in two subsets: non forbidden priority connections and non forbidden non priority or background connections.

3. Introduced VC model takes into account several parameters such as: dynamic and statistics characteristics including fractal properties of VC with feedback throughput control like TCP.

4. Considered preemptive queueing mechanism can be viewed as a background for DiffServ access control because it provides a wide range packet loss probability ratio using flexible randomized push-out algorithm.

5. Proposed push-out algorithm based on selecting priority parameter controls packet loss probability taking into account restricted capacity of packet buffer in DiffServ access point. The most interesting result obtained in congested network allows to keep priority VC throughput near the requested value, which is important for specific space experiment with robotics arm on ISS board.

REFERENCES

[1] Avrachenkov K.E., Vilchevsky N.O., and Shevljakov G.L. Priority queueing with finite buffer size and randomized push-out mechanism // Proceedings of the ACM international conference on measurement and modeling of computer (SIGMETRIC 2003). San Diego: 2003, p. 324-335.

[2] Vladimir Zaborovsky, Aleksander Gorodetsky, and Vladimir Muljukha «Internet Performance: TCP in Stochastic Network Environment», Proceedings of The First International Conference on Evolving Internet INTERNET 2009, 23-29 August 2009, Cannes/La Bocca, France, Session «INTERNET 1: Internet Performance», Published by IEEE Computer Society, 2009, p.447-452

[3] V. Zaborovsky and A. Titov «Specialized Solutions for Improvement of Firewall Performance and Conformity to Security Policy», Proceedings of The 2009 International Conference on Security and Management, Volume II, Las Vegas, Nevada, USA, July 13-16, 2009, Published by CSREA Press, USA 2009, p.603-608

[4] Zaborovsky V., Zayats O., and Muljukha V. Priority Queuing with Finite Buffer Size and Randomized Push-out Mechanism // Proceedings of the Ninth International Conference on Networks ICN 2010 Menuires, France 2010 p.316-321.

[5] http://www.dlr.de/en/desktopdefault.aspx/tabid-727

 ISBN:978-1-61208-113-7

IPv6: Now You See Me, Now You Don't

Matthew Dunlop*† Stephen Groat*† Randy Marchany† Joseph Tront*
*Bradley Department of Electrical and Computer Engineering
†Virginia Tech Information Technology Security Office
Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA
Email: {dunlop,sgroat,marchany,jgtront}@vt.edu

***Abstract*—Current implementations of the Internet Protocol version 6 (IPv6) use stateless address auto configuration (SLAAC) to assign network addresses to hosts. This technique produces a static value determined from the Media Access Control (MAC) address as the host portion, or interface identifier (IID), of the IPv6 address. Some implementations create the IID using the MAC unobscured, while others compute a onetime hash value involving the MAC. As a result, the IID of the address remains the same, regardless of the network the node accesses. This IID assignment provides third parties (whether malicious or not) with the ability to track a node's physical location by using simple tools such as ping and traceroute. Additionally, the static IID provides a means to correlate network traffic with a specific user through simple traffic analysis. We examine the techniques used to create autoconfigured addresses. We also discuss how these techniques violate a user's privacy. The serious breaches in privacy caused by SLAAC need to be addressed before deployment of IPv6 becomes widespread. To that end, we provide a detailed taxonomy of different methods for obscuring IPv6 autoconfigured IIDs.**

***Index Terms*—IPv6 addressing, privacy protection**

I. Introduction

The next generation of Internet protocol, the Internet Protocol version 6 (IPv6), implements new features based on the existing Internet Protocol version 4 (IPv4). One major change, and the driving force behind IPv6, is the address architecture. The address space in IPv4 is limited to 32 bits. Unallocated addresses in IPv4 are quickly being depleted and will be exhausted by early to mid 2011 [6], [13]. To combat the shortage of addresses in IPv4, IPv6 employs 128-bit addresses. With the current number of Internet-ready devices on the network, the immense address space provided by IPv6 is sparsely populated. However, as new classes of devices become interconnected and networked, manually managing subnets becomes complex and time consuming.

One solution being used to solve the problem of subnet management in IPv6 is stateless address auto configuration (SLAAC). SLAAC allows an administrator to configure the network and subnet portion of the address, while each device automatically configures the host portion, or interface identifier (IID), of the address. The IID is often formed by extending the 48-bit Media Access Control (MAC) address to 64 bits, spanning half of the IPv6 address.

Using a node's MAC address in the IID has serious unintended consequences to a user's privacy. While the observation that this addressing scheme could allow an attacker to analyze payload, packet size, and packet timing was made in RFC 4941 [11], the privacy implications that arise from stateless address generation have not been addressed. The issue is not only that the MAC address is used as the IID, but also that the IID remains static. As a result, no matter what network the node accesses, the IID remains the same. Consequently, simple network tools such as ping and traceroute permit tracking a node's geographic location from anywhere in the world. All the cyber-stalker needs to know is the location of the subnet.

Such "cyber-stalking" is not possible in IPv4. In IPv4, a node's MAC address is restricted to the local subnet. Additionally, a node's location is often obscured through the use of the Dynamic Host Configuration Protocol (DHCP), which leases host addresses based upon availability. Furthermore, the deployment of carrier-grade Network Address Translation (NAT) in IPv4 has the unintentional benefit of protecting a host's identity by placing it within a private address space, which is not globally addressable.

With IPv6, a user's privacy can also be violated through the monitoring of network traffic. Traffic analysis can be used to deduce an identity by correlating traffic captures from a specific IID. This analysis is possible in IPv4, but only for short periods of time, since DHCP addresses change. In contrast, a static IID permits correlation of a specific user's data over multiple sessions. The deterministic IPv6 addresses that globally tie users to each of their packets make this correlation possible. Once an attacker is able to deduce a user's identity and location, the attacker can then target the user for identity theft or other related crimes. Using static IIDs to monitor traffic for identity theft is one of many potential privacy exploits of deterministic stateless IPv6 addressing.

We will show why deterministic addresses have serious privacy implications and why this issue should be addressed before IPv6 is deployed globally. First, we provide background on IPv6 in Section II. Section III describes how the deterministic IID is formed. We discuss the privacy implications in Section IV. In Section V, we provide our taxonomy of methods for hiding users' IIDs. In Section VI, we discuss future work. We conclude in Section VII.

II. Background

The tremendous growth of the Internet has created the need for a new version of the Internet Protocol. Despite the advent of technologies such as NAT, IPv4 will soon be unable to support the addresses needed. Increasing the address space was the main motivation for developing IPv6. Researchers saw the

 ISBN:978-1-61208-113-7

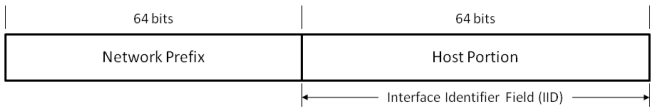


Fig. 1. IPv6 128-bit address format

need for more address space, however, as an opportunity to improve upon an already successful IPv4.

A. Benefits of IPv6

As previously discussed in Section I, IPv6 was primarily developed to support more address space in the Internet. An IPv4 address, consisting of 32 bits, provides approximately 4.2 billion possible address combinations. This address space is not sufficient to support the emerging myriad of Internet-capable devices. Therefore, the IPv6 address was expanded to 128 bits. This new address size allows for 2^{128} possible addresses, approximately $5 \cdot 10^{28}$ addresses for every one of the 6.8 billion people [17] in the world.

In addition to the larger address space, IPv6 was designed with five other main improvements. The first is simplifying the header format to 40 bytes. The second improvement makes the number of IP options extensible by moving the options out of the header and into the payload of the packet. Third, the protocol was designed to be extendible, allowing for the future definition of additional options. Flow labeling, the fourth improvement added to IPv6, allows for classification of packets belonging to particular flows. With flow labels, each router can determine which flow a packet belongs to and prioritize the packets appropriately. The fifth and final major improvement in IPv6 is the integration of authentication and encryption into the protocol stack. In IPv4, Internet Protocol security (IPsec) [9] was developed as an add-on so IPv4 did not have to be redefined. As a result, there are inefficiencies with its implementation. IPv6 solves this by integrating IPsec.

B. Stateless IPv6 Addressing

The large size of the IPv6 address space requires a new network address configuration architecture to simplify network administration. For this reason, IPv6 combines a Neighbor Discovery Protocol (NDP) [12] with SLAAC to allow for nodes to self-determine their IP addresses. Designed as a replacement for the Address Resolution Protocol (ARP), NDP facilitates nodes within a particular subnet learning of other nodes on the link using Internet Control Message Protocol version 6 (ICMPv6) messages. Once an NDP message is received, the node uses the network portion of the address to configure the first 64 bits of its IPv6 address. For the last 64 bits, the node automatically configures an address, designated as the IID of the address. The final step combines the 64-bit network address with the 64-bit host address to form a complete 128-bit IPv6 address (See Fig. 1).

NDP and SLAAC eliminate the need for DHCP addressing services currently implemented on the majority of IPv4 networks. DHCP implements a client-server architecture in which

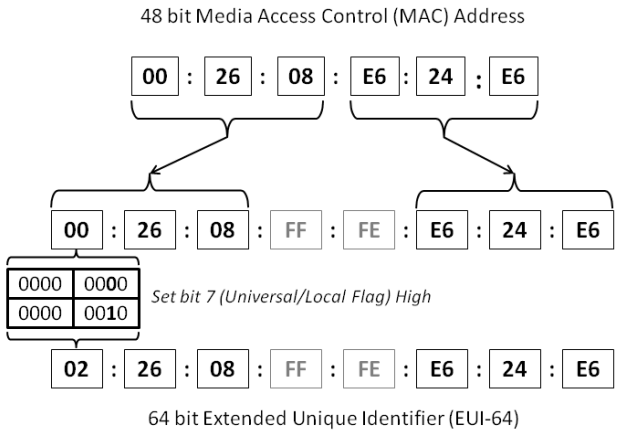


Fig. 2. 64-bit Extended Unique Identifier (EUI-64) format

a DHCP server assigns addresses to clients and keeps state of which addresses have been assigned to particular clients. DHCP also exists for IPv6 in the form of DHCPv6. The sparse address space and the ease of address autoconfiguration, however, make DHCPv6 addressing an unnecessary service. The extra expense and network complexity involved in DHCP addressing have been removed with NDP and SLAAC.

III. Deterministic IID

Due to the current accepted definition of SLAAC on most operating systems, the IID of a node's IPv6 address is deterministic across networks. For the last 64 bits, the node automatically configures an address based upon the MAC address of its network interface. By extending the 48-bit MAC address to 64 bits through the EUI-64 format [7], the IID of an IPv6 address is created. The EUI-64 format splits the 48-bit MAC address into two 24-bit halves. The 16-bit hex value `0xFFFE` is inserted between the two halves to form a 64-bit address. Also, the universal/local flag, located at bit seven of the 64-bit host portion, is set to universal. Fig. 2 illustrates this process.

While different operating systems configure IPv6 addresses differently, no current operating system implementations of IPv6 stateless addressing dynamically obscure the IID of all IPv6 addresses on the system. OS X and common Linux distributions, such as CentOS and Ubuntu, follow the EUI-64 format. The MAC address appears virtually unaltered in the IPv6 address. The Windows operating system obscures the host portion of an IPv6 address according to RFC 4941 and sets a temporary address [11], [16]. Windows operating systems, however, also carry another IPv6 address used for neighbor solicitation. This other IPv6 address contains an IID that is obscured but never changes, regardless of the subnet the node connects to. Not dynamically obscuring a user's host portion for all of the IPv6 addresses associated with a system threatens a user's privacy. The static IID currently implemented in major operating systems can be linked to a particular node, even as the node changes networks.

 ISBN:978-1-61208-113-7

Many mobile devices, such as Android and iPhone, support IPv6 in WiFi. Their implementations follow the EUI-64 format, providing these mobile devices with static IIDs that are easily tracked on their WiFi connections. Since most users frequently carry their mobile devices and leave them on and connected, the ability to track a user is increased dramatically. While the need to address the privacy concerns in Mobile IPv6 has been identified [3], [10], [14], it does little good until the privacy concerns due to IID tracking are addressed. Since Mobile IPv6 would only be applied to the cellular connections and the majority of these wireless devices also deploy WiFi, users can still be tracked through their wireless devices as they move between different WiFi networks. Therefore, address privacy must be dealt with for all connections of a mobile device to assure complete privacy.

IV. Privacy Implications

The static IID created by the EUI-64 format and the Windows operating systems compromises a user's privacy. Creating a static IID from a MAC address allows nodes to be logically and geographically tracked as they travel to different networks. Since the EUI-64 format results in a deterministic IID, users can be tracked on a network by scanning different subnets and searching for the MAC-generated IIDs. Using simple commands such as ping and traceroute, the location of a user can be determined with reasonable geographic accuracy. Even the Windows obscuration of the IID within the IPv6 address does not protect a user. By locally capturing a user's traffic once, a specific user can be paired with the deterministically obscured IID and tracked with the same technique of searching subnets as used for unobscured host addresses. Since the obscuration occurs independent of the network, a Windows host carries the same obscured IID between networks.

By monitoring the traffic on a network over an extended period of time, a single user's traffic can be identified and analyzed. Armed with this data, a third party (whether malicious or not) can potentially tie a device to its actual user. As the user crosses different subnets, traffic can be collected and correlated by examining the static IID. This vulnerability to tracking does not typically apply when using IPv4. Most medium to large IPv4 networks implement DHCP, which changes user addresses randomly. As a result, DHCP logs are needed to tie traffic sniffed from a network with a particular user. Due the to the deterministic IID in IPv6 SLAAC, simple filters could be created to filter the traffic of a single user on any subnet. This would allow an interested party to identify and monitor a user's on-line activity through traffic analysis. In a dual-stack implementation where a node uses a mixture of IPv4 and IPv6, special ICMPv6 Neighbor Solicitation messages can provide an interested party with the IPv4 address linked to an IPv6 address. This correlation allows for traffic collection to extend to IPv4 for a single session.

Tracking users or monitoring their on-line activity is not the only concern. If it was known that location or traffic monitoring was occurring, a malicious host could spoof the IID of an innocent node. The malicious node could then masquerade as the innocent node and create false traffic or locations using the innocent node's IID.

V. Possible Solutions

There are three primary classifications for protecting a user's IID from tracking and monitoring. The most straightforward classification is dynamic obscuration of the IID. The second classification is third party address assignment. The final classification is address tunneling.

A. IID obscuration

There are two methods proposed for obscuring the IID when using stateless generation of IPv6 addresses. The first method uses cryptographically generated addresses (CGAs) [2]. The second method uses privacy extensions for static IIDs [11].

1) Cryptographically Generated Addresses: CGAs are IPv6 IIDs that are generated by hashing a sender's public key and other parameters [2]. Although CGAs can be used with multiple applications, they were originally designed to work with the SEcure Neighbor Discovery (SEND) protocol [1] to prevent denial of service attacks. Since CGAs obscure a sender's IID, they could prevent tracking a user through their autoconfigured IPv6 address. There are, however, a number of disadvantages to using CGAs for dynamically obscuring an IPv6 IID.

The main disadvantage to using CGAs is the cryptographic cost of CGA generation. In order to generate a CGA, the sender must apply the SHA-1 hash algorithm to the CGA parameters a minimum of two times. The CGA parameters consist of the concatenation of a pseudo-random number, the sender's public key, nine zero octets, and optional fields. The first hash calculation generates Hash2. The ease of achieving an acceptable Hash2 value depends on the strength of the security parameter (Sec). The Sec is a three-bit value used to determine how many leading zeros Hash2 must contain. Hash2 must have $16 \cdot Sec$ leading zeros. If it does not, a pseudo-random number is incremented and Hash2 is recalculated. Hash2 continues to be recalculated until this condition is met. This can result in a large number of hash calculations. In fact, RFC 3972 states that on average it takes O($2^{16 \cdot Sec}$) iterations to generate a CGA [2].

After Hash2 is successfully calculated, Hash1 is calculated using the CGA parameters with the final pseudo-random number used to generate Hash2. The left-most 64 bits of the SHA-1 output are used as the IID with the exception of bits 0-2, 6 and 7 used for other purposes. Duplicate address detection is conducted as required [1]. If a collision occurs, a collision count field is incremented and Hash1 is recalculated. This is repeated at most three times after which, CGA calculation stops [2]. It is possible that after all of these hash calculations no CGA will result and the process will need to start anew.

These repeated hash calculations require minimal overhead for the average personal computer, but are likely infeasible for most handheld devices due to limited power and computational capability. To help mitigate this cost, RFC 3972 discusses that CGAs can be precomputed or offloaded to more powerful

 ISBN:978-1-61208-113-7

computers [2]. This solution, however, does not help a user connecting to a network for the first time, which is not uncommon when roaming with a handheld device. Additionally, the CGA generation cost makes it prohibitive for users to generate new CGAs. Thus, CGAs for subnets will be deterministic.

There are two possible implementation alternatives to reduce the cost of CGA generation. The first involves setting $Sec = 0$. By setting $Sec = 0$, Hash2 is not required to have any leading zeros. This means that the calculation of Hash2 is not required [2]. The problem with this alternative is that the CGA becomes susceptible to a brute-force attack. If the CGA is used only to prevent IPv6 address tracking, this may not be a concern. However, if the intent is to also protect the IID from denial of service or address spoofing, this alternative is not acceptable.

The second implementation alternative is to omit the subnet prefix from CGA generation [2]. By omitting the subnet prefix, senders need only calculate the CGA once. The same CGA can be used for multiple subnets without having to go through the expensive CGA generation for every subnet. The downside is that the IID is again deterministic across multiple subnets. This alternative maintains the benefit of protection against brute-force attack through the security parameter, but leaves senders vulnerable to address tracking.

Even if users do change their CGA every time they connect to a subnet, they can still be tracked. Senders using CGAs send their CGA parameters along with their IPv6 address for verification purposes. The CGA parameters contain the users' public key. Unless senders also generate new public keys for every connection, they can be tracked through their CGA parameters.

An additional disadvantage of using CGAs is that an attacker can easily impersonate or slander a user by forming a CGA claiming to be the target. Asymmetric key pairs are self-generated to eliminate the key management infrastructure. As a result, anyone can generate a key pair and claim to be someone else. RFC 3972 claims that CGAs protect against address spoofing [2]. This is only true in the case where an attacker attempts to hijack an existing session between two nodes. An attacker is free to initiate new sessions claiming to be someone else. By doing this, an attacker can make it appear as if a target is behaving in a certain way to others who are monitoring the target on the network. Spoofing is common in IPv4, but CGAs in IPv6 make a spoofed address appear more authentic.

Requiring the use of public keys to form CGAs is another drawback. One of the strengths of stateless autoconfiguration of IPv6 addresses is that it is transparent to users. Requiring users to generate asynchronous key pairs removes this transparency. Many users do not have asynchronous key pairs or even understand the concept. As a result, most users will not implement CGAs to obscure their addresses.

The remaining vulnerabilities do not affect tracking of users, but how an attacker can take advantage of flaws in CGAs to achieve malicious results. For instance, a man-in-the-middle can cause all data checks to fail by modifying any of the CGA parameters sent with the IPv6 address. This can be accomplished by changing the collision count to a number greater than two, because the collision count is only valid for values of 0-2. An attacker also can hijack a connection by finding a CGA hash collision. Only 59 bits (64 bits minus bits 0-2, 6, and 7) of the SHA-1 hash are used for the CGA. Using the birthday attack [15], a collision can be found in $2^{59/2}$ guesses on average.

Despite the numerous disadvantages of CGAs, there are several advantages. The primary advantage is, of course, that CGAs obscure a sender's IID. Even if senders do not go through the costly procedure of computing a new CGA every time they connect to a network, they will at least have a different deterministic CGA for every subnet. This will make an attacker's attempt at tracking a target much more difficult. In order for an attacker to track a target, the attacker will need to know the sender's CGA on every subnet. Another benefit of using CGAs is that they require no key management infrastructure [2]. Since CGAs can use user-generated asymmetric key pairs, there is no need for a certification authority, making CGAs scalable.

2) Privacy Extensions: The privacy extensions proposed in RFC 4941 provide another way to obscure the IPv6 IID. Where the goal of CGAs is primarily to provide security of IPv6 addresses for the SEND protocol as discussed in Section V-A1, the goal of implementing privacy extensions is to prevent an attacker from correlating network traffic with a particular user. Privacy extensions produce a random IID by hashing the concatenation of a user's IID and a 64-bit "history value." This "history value" is initially generated from the rightmost 64 bits of a hashed random number. Subsequently, the "history value" becomes the previously calculated 64-bit hash result. The random IID is formed using the leftmost 64 bits of the resultant hash calculation with the "u" bit (bit 6) set to zero for local scope. Duplicate address detection is then performed on the random IID to detect duplicate IIDs on the local network. If a duplicate IID is detected, a new "history value" is formed and the process is repeated [11].

Using privacy extensions to obscure a user's IID is much more appealing than using CGAs for several reasons. First, the cryptographic cost is much lower. Assuming no duplicate address is detected, privacy extensions require only one hash calculation by the sender at IID generation and none by the receiver. With CGAs, the sender requires on average $O(2^{16 \cdot Sec})$ hash calculations to generate a CGA, and the receiver is required to complete two hashes to verify the CGA. The cryptographic cost of using privacy extensions is more feasible, allowing a node to change their IID often.

The use of privacy extensions does not require the use of a public key [11]. As a result, implementation of privacy extensions are more transparent to the user. Additionally, there are no required accompanying parameters that can be used to link a sender to the random IID.

Privacy extensions are more effective at obscuring IIDs than CGAs because of the lifetime of a generated IID. Having users generate new IIDs when they connect to new subnets

 ISBN:978-1-61208-113-7

effectively masks users' activities on the Internet, but it does nothing for users that never migrate from a particular subnet. A CGA user that never moves between subnets (e.g., a desktop computer) would not necessarily ever generate a new random IID. The privacy extensions specification uses a TEMP_VALID_LIFETIME parameter to set the maximum amount of time a random IID can be used before needing to be regenerated [11]. This feature improves on DHCP addresses that may last months or longer.

Implementing random IIDs through privacy extensions does have disadvantages. Just as with CGAs, IID collisions are possible [11]. Since only 64 bits of the hash calculation are used, the chance of hash collisions increases. Similar to CGAs, the generation process terminates after a set number of IID collisions.

As mentioned previously, privacy extensions specify several different parameters to limit the time a random IID is valid. The default values of these parameters are set too long. The TEMP_PREFERRED_LIFETIME parameter is set to one day and specifies the preferred life of a random IID. The default value of TEMP_VALID_LIFETIME is one week. An application can force an address to use TEMP_VALID_LIFETIME rather than TEMP_PREFERRED_LIFETIME. Additionally, RFC 4941 states that a new random IID *should* be computed when connecting to a new subnet [11]. It does not state that this *must* be done. Users choosing not to change IIDs and accepting the default expiration times could be monitored for up to a week at a time before an attacker needs to reestablish an identity pairing. Fortunately, RFC 4941 allows users to modify these defaults.

An unfortunate side-effect of privacy extensions is that IIDs are obscured from everyone, including network administrators. Combined with frequently changing and obscured IIDs, privacy extensions make fault isolation and debugging difficult [11]. With CGAs, public keys are tied to addresses. When DHCP is used, system administrators can track changes in addresses, while still protecting a user's identity.

CGAs are more robust than privacy extensions in that they can protect a sender's address from spoofing once a session has been initiated. Since there is no verification process implemented by privacy extensions, an attacker can easily inject new traffic claiming to be the sender. However, preventing this type of malicious activity is not one of the design goals of privacy extensions. This type of attack, however, is mitigated by frequently changing a user's IID.

B. Third Party Address Assignment

DHCPv6 provides third party address assignment for IPv6 [5]. Instead of allowing a client to configure his/her own address, a DHCP addressed network requires a DHCP server in order for a client to get an address on a network. DHCP addresses are leased to clients when they connect to the network. If the network is not overloaded with clients, a client may receive the same DHCP address each time he/she connects. For heavily populated networks, clients may receive different addresses each time they connect.

There are several advantages to using DHCPv6. First, DHCPv6 provides IIDs that are not necessarily tied to a user. A user may receive a different DHCPv6 address each time he/she connects. Whereas collisions were possible with the address obscuration techniques discussed in Section V-A, a properly configured DHCPv6 server should not issue duplicate addresses. Although DHCPv6 can operate without any cryptographic cost for address generation or verification, it is worth noting that DHCPv6 can be configured to use optional authentication [5]. DHCPv6 authentication uses shared keys which, although more cryptographically efficient, do not scale well.

Despite these advantages, DHCPv6 does not necessarily protect a user's identity well because the IPv6 address space is sparsely populated. As a result, there will likely be little competition for addresses when connecting to even the most populated subnets. Therefore, users should expect to get the same address each time they connect to the network. This may change over time as more devices connect to the network, and the address space becomes more densely populated. However, the DHCPv6 specification promotes static addresses. Unless specifically requested by the client using the Identity Association for Temporary Addresses (IA_TA) option, the client will be issued a non-temporary address [5]. A static address, whether generated by the client itself or issued by a DHCP server, exposes the user to monitoring. If IA_TA is implemented, it would be provided through the use of privacy extensions discussed in Section V-A2.

Even if temporary addresses are requested, an attacker may still be able to monitor a user. For example, the server controls how often a client's address changes, not the client. The client does have the option to specify a preferred-lifetime and a valid-lifetime, but the timing of an address change is ultimately determined by the server. Additionally, communications between client and server use what is called a DHCP Unique Identifier (DUID). The DUID is a globally unique value that should never change [5]. The DUID is used in many of the DHCP exchange messages and could be used by an attacker to track a target's presence on a network. Once the attacker locates the user, the attacker can monitor the DHCP exchanges to harvest the IPv6 address. The scope of this attack is limited to the subnet of the user, the DHCP server, or any relays.

A final downside to using DHCPv6 versus stateless address autoconfiguration is that it must be managed. This could add a tremendous burden on network administrators as more and more devices connect to the network. DHCPv6 also increases the chances of incorrect configuration, which may lead to other vulnerabilities.

C. Address Tunneling

Address Tunneling can be achieved through the use of Internet Protocol Security (IPsec). IPsec is not a very comprehensive method to use for masking IPv6 addresses, but it can provide excellent obscuration from external attackers. IPsec provides authentication and/or encryption to network layer packets. For the purposes of obscuring IPv6 addresses, we re-

 ISBN:978-1-61208-113-7

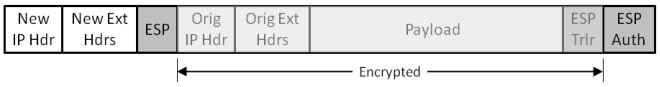


Fig. 3. IPv6 packet encrypted using IPsec in tunnel mode

fer only to the encryption aspect provided by the encapsulating security payload (ESP) as illustrated in Figure 3. Specifically, ESP used in tunnel mode provides address obscuration. When used in tunnel mode, ESP encrypts the entire IPv6 packet and provides a new IPv6 header, complete with new source and destination addresses [8]. The new source and destination addresses are those of the endpoints of the tunnel. It should not be possible for devices external to the tunnel to learn the true identity of the sender or receiver. The tunnel endpoints are typically network gateway devices. Although it is possible for the sender and receiver to act as tunnel endpoints in IPv6, this technique would gain very little privacy. A host acting as its own tunnel endpoint would be easy to link as the actual target host. The ability to link to the target host can be prevented by using one of the obscuration techniques described in Section V-A after applying ESP. Using IPsec for this purpose, however, would then be pointless.

There are three main benefits to using IPsec in tunnel mode to obscure IPv6 addresses. As mentioned previously, the sender's address is hidden from those external to the tunnel. Second, the cryptographic burden of encryption and decryption is offloaded to the gateway devices. This makes address obscuration feasible for devices limited by battery or computational power, such as handheld devices and sensors. Also, since the address used is that of the gateway devices, there are no address collisions as there were in the address obscuration techniques discussed in Section V-A.

Unfortunately, IPsec in tunnel mode does not provide any address protection from an attacker inside the sender's or receiver's subnets. Since address obscuration does not occur until the packet reaches the gateway, an attacker monitoring the subnet will have no trouble monitoring users through their IIDs. This does, however, limit the scope of the attack to the two subnets mentioned.

Perhaps an even bigger issue is the requirement for a key management infrastructure. Wide-scale deployment of IPsec requires a global trust model in place as well as a management infrastructure [4]. In today's infrastructure, only those networks with security requirements utilize IPsec. It is not reasonable to assume that networks would implement IPsec for address obscuration purposes.

VI. Future Work

The next phase of our research is to design and implement an address obscuration technique. Each of the techniques described in Section V has associated shortcomings. Those that obscure the IPv6 IID, are not feasible for resource constrained devices. Third-party address assignment and address tunneling have scope limitations. Our technique is designed to minimize computational complexity while avoiding scope limitations, thus being feasible for power-constrained devices. We also plan to test the validity and overhead of our design using our campus-wide IPv6 production network.

VII. Conclusion

IPv6 is a definite improvement over IPv4, allowing more devices to connect to the Internet using globally unique addresses. SLAAC, however, violates a user's privacy and needs to be addressed before IPv6 is deployed. A number of different methods can be used to obscure a user's IID from monitoring. Each method comes with its associated benefits and shortcomings. Currently, the privacy extensions outlined in RFC 4941 [11] appear to provide the best balance of both. Privacy extensions provide an unmanaged solution to address obscuration with low cryptographic cost. However, the current lack of computational power of many handheld devices combined with possible address collisions are likely obstacles to implementing this algorithm. Regardless of which solution is implemented, some method of obscuring IIDs should be deployed as part of operating systems and embedded devices to protect the privacy of users.

References

[1] J. Arkko, J. Kempf, B. Zill, and P. Nikander. SEcure Neighbor Discovery (SEND). RFC 3971 (Proposed Standard), Mar. 2005.

[2] T. Aura. Cryptographically Generated Addresses (CGA). RFC 3972 (Proposed Standard), Mar. 2005. Updated by RFCs 4581, 4982.

[3] C. Castelluccia, F. Dupont, and G. Montenegro. A simple privacy extension for mobile IPv6. In *Mobile and Wireless Communication Networks, IFIP TC6 / WG6.8 Conference on Mobile and Wireless Communication Networks (MWCN 2004)*, pages 239–249, Oct. 2004.

[4] A. Choudhary. In-depth analysis of IPv6 security posture. In *The 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2009)*, pages 1 –7, Nov. 2009.

[5] R. Droms, J. Bound, B. Volz, T. Lemon, C. Perkins, and M. Carney. Dynamic Host Configuration Protocol for IPv6 (DHCPv6). RFC 3315 (Proposed Standard), July 2003. Updated by RFCs 4361, 5494.

[6] A. Gonsalves. IP addresses predicted to be exhausted in 2011. *InformationWeek*, July 2010.

[7] R. Hinden and S. Deering. IP Version 6 Addressing Architecture. RFC 4291 (Draft Standard), Feb. 2006.

[8] S. Kent. IP Encapsulating Security Payload (ESP). RFC 4303 (Proposed Standard), Dec. 2005.

[9] S. Kent and K. Seo. Security Architecture for the Internet Protocol. RFC 4301 (Proposed Standard), Dec. 2005.

[10] R. Koodli. IP Address Location Privacy and Mobile IPv6: Problem Statement. RFC 4882 (Informational), May 2007.

[11] T. Narten, R. Draves, and S. Krishnan. Privacy Extensions for Stateless Address Autoconfiguration in IPv6. RFC 4941 (Draft Standard), Sept. 2007.

[12] T. Narten, E. Nordmark, W. Simpson, and H. Soliman. Neighbor Discovery for IP version 6 (IPv6). RFC 4861 (Draft Standard), Sept. 2007.

[13] Remaining IPv4 address space drops below 5%. Available at: http://www.nro.net/media/remaining-ipv4-address-below-5.html/, Oct. 2010.

[14] Y. Qiu, J. Zhou, F. Bao, and R. Deng. Protocol for hiding movement of mobile nodes in Mobile IPv6. In *62nd IEEE Vehicular Technology Conference*, volume 2, pages 812 – 815, Sept. 2005.

[15] W. Stallings. *Cryptography and network security (2nd ed.): principles and practice*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[16] Introduction to IP version 6. Available at: http://download.microsoft.com/download/e/9/b/e9bd20d3-cc8d-4162-aa60-3aa3abc2b2e9/ipv6.doc accessed on 24 May 2010.

[17] U.S. & World population clocks. Available at: http://www.census.gov/population/www/popclockus.html/ accessed on 4 Mar 2010.

 ISBN:978-1-61208-113-7

A Review Study on Image Digital Watermarking

Charles Way Hun Fung
CPGEI / UTFPR
Avenida Sete de Setembro, 3165
Curitiba-PR - CEP 80.230-910
E-mail: charleswhfung@gmail.com

Antônio Gortan
CPGEI / UTFPR
Avenida Sete de Setembro, 3165
Curitiba-PR - CEP 80.230-910
E-mail: gortan@dgmdesign.com.br

Walter Godoy Junior
CPGEI / UTFPR
Avenida Sete de Setembro, 3165
Curitiba-PR - CEP 80.230-910
E-mail: godoy@utfpr.edu.br

***Abstract*—There has been an increase in broadcasting media since the begin of this century, because many techniques had been developed to solve this problem. Watermarking is the greatest bet from many researchs around the world. Digital watermarks can be used by a lot of applications like: copyright protection, broadcast monitoring and owner identification. In this paper, we will show a classification of watermarks, propose a basic model for watermarking and explain some recent algorithms for image watermarking and their features, citing examples applicable to each category.**

***Index Terms*—digital watermarking; wavelets; security; dwt; lwt; svd.**

I. INTRODUCTION

The increased Internet usage has turned a technique that is able to protect the copyright of published medias into a necessity. The easy of distribution of these documents through the web may transgress protection laws against unauthorized copies and make fidelity questionable. Digital watermarking has been proposed as a solution against these practices.

Digital watermark is a labeling technique of digital data with secret information that can be extracted in the receptor. The image in which this data is inserted is called cover image or host [1]. The watermarking process has to be resilient against possible attacks, keeping the content of the watermark readable in order to be recognized when extracted. Features like robustness and fidelity are essentials of a watermarking system, however the size of the embedded information has to be considered since data becomes less robust as its size increases. Therefore a trade-off [2] of these features must be considered.

The paper is organized as follows. In Section II, we described the classification and each feature. In Section III, we explain the main applications of watermarking. A basic model and the discussion about each block of the process that is proposed in Section IV. Section V is the conclusion and Section VI the acknowledgments.

II. CLASSIFICATION

A watermarking system has requirements which must be met when implemented, however the application will dictate which features should be emphasized. In this section a classification of marks according with their requirements will be proposed.

This paper is financial supported by CAPES-Brazil.

A. Robustness:

This feature refers to the ability to detect the watermark after some signal processing operation [1]. Marks cannot survive all kinds of attacks, hence attacks resilience must be optimized according to application. For example: To verify data integrity a correlation between the received image and the signal is carried out when the watermark is extracted. If differences are found then manipulations must have occured [3]. With that in mind the following classification can be made:

1) Fragile: These marks can be destructed by small manipulations of the watermarked image [4]. Such marks have been used for authentication and integrity verification.

2) Semi Fragile: These behave as fragile watermarks against intentional modifications and as robust watermarks against casual manipulations [5] like noise. These marks have been used in image authentication and tamper control.

3) Robust: According to [4], these watermarks are designed to resist heterogeneous manipulations. They can be used in copy control e monitoring.

B. Fidelity:

This requirement could be called invisibility. It preserves the similarity between the watermarked object and the original image according to human perception [1]. The mark must remain invisible notwithstanding the occurrence of small degradations in image brightness or contrast.

C. Capacity or Data Payload:

The number of bits that can be inserted through watermarking varies with each application. In case of images, a mark will be a static set of bits. In videos, capacity will be gauged by the quantity of inserted bits per frame, in audio files by the quantity of inserted bits per second [1].

D. Detection Types:

This classification determines which resources are necessary for the analysis to extract the watermark from the cover image.

1) Blind: In this detection type the original image and mark data is not available to the receiver. For example: Copy control applications must send different watermarks for each user and the receiver must be able to recognize and interpret these different marks [1].

 ISBN:978-1-61208-113-7

2) Non-Blind: In this case, the receiver needs the original data, or some derived information form it, for the detection process [1]. This data will also be used in the extraction algorithm.

E. Embedding:

The method used to embed the watermark influence both the robustness against attacks and the detection algorithm, but some methods are very simple and cannot meet the application requirements. El-Gayyar and von zur Gathen [2] showed that designing a watermark should consider a trade-off among the basic features of robustness, fidelity and payload.

There are two approaches for the embedding process:

1) Spatial Domain: These watermarks insert data in the cover image changing pixels or image characteristics [4]. The algorithms should carefully weight the number of changed bits in the pixels against the possibility of the watermark becoming visible [2]. These watermarks have been used for document authentication and tamper detection.

2) Transform Domain: These algorithms hide the watermarking data in transform coefficients, therefore spreading the data through the frequency spectrum [1] making it hard to detect and strong against many types of signal processing manipulations. The most used transforms are: Discrete cosine transform (DCT) [1], discrete wavelet transform (DWT) [6] and discrete lifting transform (LWT) [7].

III. Applications

Before discussing watermarking algorithms let us review some common applications:

A. Broadcast Monitoring:

This type of monitoring is used to confirm the content that is supposed to be transmitted [1], [3], [8]. As an example, commercial advertisements could be monitored through their watermarks to confirm timing and count.

B. Owner Identification:

The conventional form of intellectual ownership verification is a visual mark. But, nowadays, this is easily overcome by the use of softwares that modify images. An example is images with a copyright registration symbol © which have this mark removed by specialized softwares. In this case invisible watermarks are used in order to overcame the problem.

C. Fingerprinting:

A watermarked object contains information about the owner permissions. Several fingerprints can be hosted in the same image since the object could belong to several users [3], [8].

D. Publication monitoring and copy control:

The watermark contains owner data and specifies the corresponding amount of copies allowed. This presupposes a hardware and a software able to update the watermark at every use [3]. It also allows copy tracking of unauthorized distribution since owner data is recorded in the watermark.

IV. Basic Model

Liu and He [3] present a model with three stages: Generation & Embedding, Distribution & Possible Attacks and Detection. In this paper, we adapted this model dividing the first block in Generation and Embedding, because the both use different watermarking algorithms and can be studied independently. The basic model proposed is presented below:

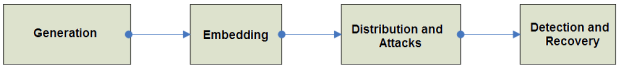


Fig. 1. Basic Model

The explanation about the Embedding and Detection stage will be presented together, because the algorithms are related. In Fig 1 the basic model can be divided in four stages:

A. Generation

In this stage the mark is created and its contents must be unique and complex in order to be difficult to extract or not to be damaged by possible attacks. Some algorithms that have been used for watermark generation will presented below:

1) Images in grayscale or binary: Many marking objects can be images or brands that represent some enterprise or have some data that identify the cover image. Depending on the application the marks can be binary images or grayscale containing a larger amount of data and even some intrinsic features that helps in the extraction process [8].

2) Pseudo-Random Sequence: This mark has a random seed that is used to generate the marked matrix. This seed must be stored like a secret key and will be used in the detection process to reconstruct the mark. The use of binary marks with this algorithm is rather common.

3) Chaotic Sequence: The watermarks are prepared using maps of chaotic functions [9]–[11]. These sequences are easy to implement because there are predefined models to create them. Due to statistic features these watermarks resist several types of attacks, like simple attacks and distortion.

4) Error Correcting Codes and Cryptography: The insertion of redundancy in watermarks or in the cover image can improve the extraction process or the reconstruction of the watermark after attacks. However codes can cause collateral effects by increasing the amount of embedding data, which may in turn harm the watermark robustness or decrease its data payload. The most commonly used codes are: Hamming [12] Bose-Chaudhuri-Hocquenghen (BCH) [12]–[14], Reed Solomon [12], [13], Low density parity check (LDPC) [15], and Turbo [16].

B. Embedding and Detection:

The embedding is directly related with the extraction algorithm, in this section we will discuss how this has been done in recent algorithms. The embedding algorithm is basically a combination of the watermark with the chosen media [3], so the result is equivalent to:

 ISBN:978-1-61208-113-7

$$I_W = E(I, W) \quad (1)$$

where I is the original media, W the watermark, E is the embedding function and I_W the watermarked media. The function depends on the algorithm and the analyzed domain.

1) Spatial Domain: In this case the embedded watermark is equivalent to noise addition to the original media, thereby influencing the watermarked object characteristics. Two following we will be presented below:

- **Least Significant Bit (LSB)**:

This is the simplest approach, because the least significant bits carry less relevant information and their modification does not cause perceptible changes. Among these approaches there are types using only the salient points [17] or type, which use some kind of cryptography on the watermark message before the embedding process [18], In this last case, a cipher called "datamark" is created, which is embedded in the cover image using a key. This key determines which points must be modified by the embedding process.

The extraction algorithm is the inverse of embedding. The marked object must be analyzed and its least significant pixel bits isolated. These extracted bits can be used together with the cryptography keys in decoding algorithms to recover the original watermark.

- **Singular Value Decomposition (SVD)**:

It is a numeric analysis of linear algebra which is used in many applications in image processing. It is used to decompose a matrix with a little truncate error according to the equation below:

$$A = USV^T \quad (2)$$

Where A is the original matrix, U and V are orthogonal matrices with dimensions MxM and NxN respectively, S is a diagonal matrix of the Eigenvalues of A and T indicates matrix transposition. [19] did the decomposition of the cover image and added the watermark using a scale coefficient α to get the following equation:

$$S + \alpha W = U_W S_W V_W^T \quad (3)$$

Multiplying matrices U, V^T and S_W result in the marked image A_W:

$$A_W = US_W V^T \quad (4)$$

This was possible due to the high stability of singular values (SV) of SVD. In another approach, the cover image was separated in blocks and the SVD applied to each block [20], in this case the dimension of watermark must be equal to the blocks size and a copy of the watermark is embeded in each block. This method improves watermark robustness and resistance against many kinds of attacks.

2) Transform Domain: The mark is embedded into the cover image spectrum, thus not directly influencing the selected image quality. The following transforms are used, among others, in image spectral analysis: DCT, DWT. Some watermarking algorithms using these transforms are presented below:

- **Discrete Cosine Transform (DCT)**:

The DCT makes a spectral analysis of the signal and orders the spectral regions from high to low energy. It can be applied globally or in blocks. When applied globally, the transform is applied to all parts of the image, separating the spectral regions according to their energy. When applied in blocks, the process is analogous, only the transform is applied to each block separately.

Below, we list the typical algorithm steps found in the literature [1], [8]:

1) Segment the image into non-overlapping blocks of 8x8;
2) Apply forward DCT to each of these blocks;
3) Apply some block selection criteria;
4) Apply coefficient selection criteria;
5) Embed watermark by modifying the selected coefficients;
6) Apply inverse DCT transform on each block.

- **Discrete Wavelet Transform (DWT)**:

The wavelet transform decompose the image in four channels (LL, HL, LH and HH) with the same bandwidth thus creating a multi-resolution perspective. The advantage of wavelet transforms is to allow for dual analyses taking into account both frequency and spatial domains.

Wavelets are being widely studied due to their application in image compression, owing to which compression resistant watermarks may be achieved through their use. Another interesting feature of the DWT is the possibility to select among different types of filter banks, tuning for the desired bandwidth. The most commonly used filters are: Haar, Daubechies, Coiflets, Biorthogonal, Gaussian.

When the DWT is applied to an image, the resolution is reduced by a 2^K, where K is the number of times the transform was applied.

These algorithms are called the "Wavelet based Watermarking" [8]. The watermark is inserted by substituting the coefficients of the cover image for the watermark's data. This process improves mark robustness, but depends on the frequency. The low frequency (LL) channel houses image contents in which a coefficients change, however small, will damage the cover image, which in turn challenges the fidelity propriety. However when this region of the spectrum is watermarked, a robust mark against compressions like JPEG and JPEG2000 is attained. Furthermore, when the middle and high frequency channels are marked, some benefits against noise interference and several types of filtering show up. Therefore these algorithms tend to be adapted for human visual system (HSV) to avoid small modification in the cover image being perceptible.

 ISBN:978-1-61208-113-7

Taskovski et al. [21] implemented two watermarks using binary marks in LL2 and HH2 respectively, resulting in a mark which is robust against manipulations like compression and weak against cropping and rescaling. Similarly, [22] created a watermark adapted to JPEG2000 using two algorithms to modify the wavelet coefficients of the LH2 band of the cover image, introducing only minimal differences between the watermarked image and the original. The decision, which algorithm to use, is based on which one produces the smallest change.

To create a watermark which is resistant against noise and some kinds of processing [23] proposed an algorithm that makes three watermarks: pseudo-random, luminance and texture. The first mark is embedded in LL1 band and the others are inserted by segmenting the cover image in blocks and ordering according to the sum of coefficients and standard deviation. This algorithm is robust against cropping, noise and several compression levels.

In order to increase its recovery capacity, error correcting codes can be applied to the watermark; however, its storage capacity will be reduced due to the additional redundancy. A performance comparison of the Hamming, BCD, and Reed-Solomon codes is presented in [12]. For small error rates, the codes are effective in error elimination when compared to no coding; on the other hand for higher rates, no benefit has been observed.

Mixing spatial and transform analysis, we have a robust watermark with different features. An algorithm that applies the SVD in all bands of the first level of DWT is proposed [24], making this a watermarking process in all frequencies. Bao [25] made a watermark of the singular values (SV) of each band of the cover image, in order to achieve the least possible distortion according to the human visual system. This watermark is resistant against JPEG encoding, but is fragile against filter manipulation and random noises. An algorithm with greater robustness against cropping, Gaussian noise and compression is proposed in [24]. Initially, the DWT is applied to HL1 or HH1. In the selected band, HH2 or HL2 must be selected and divided into 4x4 blocks. Finally, SVD is applied to each block, and the watermark is embedded into the S matrix.

- **Lifting Wavelet Transform (LWT)**:

Also called second generation wavelet transforms, its use has grown due to low memory consumption and easy implementation [26]. The following LWT scheme below is adapted from [7]:

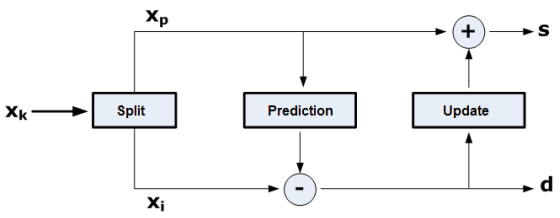


Fig. 2. Lifting scheme

In Figure 2, there are three basic operations: split, predict and update.

In split stage the input x_k is separated into odd (x_i) and even (x_p) samples, so that each of these variables contains half the number of samples of x_k.

In the prediction stage, even samples are used to predict the odd samples. The details coefficients or high frequency (h) are calculated as prediction errors of the odd samples through the use of the prediction operator P:

$$h = x_i - P(x_p) \tag{5}$$

To create the low frequency samples s, the even samples are updated through the update operator U:

$$s = x_p - U(d) \tag{6}$$

Some approaches use the LWT in spatial domain operations, as in [27], which embedded a watermark in band LL3, changing the least significant bits of the wavelet coefficients. On the other hand, [26] used a combination of SVD and LWT to apply two levels of wavelet to the cover image and select among one of bands: HH2, HL2 and LH2. In that approach, SVD is applied separately to the watermark and the selected band. The resulting S matrices must be combined into an S matrix, which will be used to create the watermarked image. This process is not blind, however it is robust against many types of manipulations like: noises, rotation, JPEG compression and quantization. It also exhibits very good performance concerning PSNR and normalized correlation values.

C. Distribution and Attacks:

The transmission media can cause some loss in the signal implying in a damaged content. These attacks may be intentional or accidental [3]. Intentional attacks use all available resources to destroy or modify the watermark making it impossible to extract it, the methods usually used are: signal processing techniques, cryptanalysis, steganalysis. On the other hand, accidental attacks are inevitable, because every image processing or transmission noise may introduce distortions.

Hartung et al. [28] classified these attacks in classes:

1) Simple Attacks: These attacks change the data of the cover image without attempting to target the watermark location. Example: Noise addition, cropping, conversion to analog and wavelet-based compression.

2) Disabling Attacks: The goal of these attacks is to attempt to break the correlation between the watermark and the cover image, making extraction impossible. Example: Geometric distortions, rotation, cropping and insertion of pixels.

3) Ambiguity Attacks: These attacks confuse the receptor embedding a fake watermark, making it impossible to discover which was the original embedded mark in the cover image.

4) Removal Attacks: In this type of attack a study of the watermark is carried out, estimating the watermark content and attempting to separate it from the host image. Example: Certain non-linear filter operations and attacks tailored to a specific watermark algorithm.

V. Conclusion

In this paper, we have reviewed some recent algorithms, proposed a classification based on their intrinsic features, embedding methods and detection forms. Also a basic four steps model for the watermark process was presented.

Many watermarking algorithms have been reviewed in the literature which show advantages in systems using wavelet transforms with SVD. These marks are robust against several different attacks. Another highlight is the replacement of DWT by LWT which improves computational performance and has an easier hardware implementation.

In future works, the use of coding and cryptography watermarks will be approached. There is a large amount of literature on these topics showing that robustness increments can be gained through the addition of coding techniques.

Acknowledgments

The authors gratefully acknowledge CAPES-Brazil for financial support.

References

[1] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*. Morgan Kaufmann, 2008.

[2] M. El-Gayyar and J. von zur Gathen, "Watermarking techniques spatial domain," University of Bonn Germany, Tech. Rep., 2006.

[3] J. Liu and X. He, "A review study on digital watermarking," *First International Conference on Information and Communication Technologies*, pp. 337–341, 2005.

[4] M. Arnold, M. Schmucker, and S. D. Wolthusen, *Techniques and Applications of Digital Watermark and Content Protection*. Artech House, 2003.

[5] X. Wu, J. Hu, Z. Gu, and J. Huang, "A secure semi-fragile watermarking for image authentication based on integer wavelet transform with parameters," *Australasian Information Security Workshop*, vol. 44, 2005.

[6] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transaction on Image Processing*, vol. 1, pp. 205–220, 1992.

[7] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *SIAM Journal on Mathematical Analysis*, 1997.

[8] V. M. Potdar, S. Han, and E. Chang, "A survey of digital image watermarking techniques," *3rd IEEE International Conference on Industrial Informatics*, pp. 709–716, 2005.

[9] A. Tefas, A. Nikolaidis, N. Nikolaidis, V. Solachidis, S. Tsekeridou, and I. Pitas, "Performance analysis of watermarking schemes based on skewtent chaotic sequences," *NSIP'01*, 2001.

[10] S.Teskeridou, V. Solochidis, N. Nikolaidis, A. Tefas, and I. Pitas, "Bernoulli shift generated chaoticwatermarks: Theoretic investigation," *SCIA2001*, 2001.

[11] W. Yan, Z. Shi-qiang, and W. Yan-chun, "Wavelet digital watermark based on chaotic sequence," *ICICIC'08*, 2008.

[12] K. L. W. G. Natasa Terzija, Markus Repges, "Digital image watermarking using discrete wavelet transform: Performance comparison of error correction codes," *Visualization, Imaging and Image Processing*, 2002.

[13] L. Haiyan, Z. Xuefeng, and W. Ying, "Analysis of the performance of error correcting coding in audio watermarking," *3rd IEEE Conference on Industrial Electronics and Applications*, pp. 843–848, 2008.

[14] P. Cika, "Watermarking scheme based on discrete wavelet transform and error-correction codes," *16th International Conference on Systems, Signals and Image Processing*, pp. 1–4, 2009.

[15] A. Bastug and B. Sankur, "Improving the payload of watermarking channels via ldpc coding," *Signal Processing Letters*, vol. 11, pp. 90–92, 2004.

[16] C. Nafornita, A. Isar, and M. Kovaci, "Increasing watermarking robustness using turbo codes," *International Symposium on Intelligent Signal Processing*, pp. 113–118, 2009.

[17] N. Pantuwong and N. Chotikakamthorn, "Line watermark embedding method for affine transformed images," *ISSPA 2007*, pp. 1–4, 2007.

[18] S. Riaz, M. Y. Javed, and M. A. Anjum, "Invisible watermarking schemes in spatial and frequency domains," *International Conference on Emerging Technologies*, 2008.

[19] R. Liu and T. Tan, "An svd-based watermarking scheme for protecting rightful ownership," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 4, pp. 121–128, 2002.

[20] R. A. Ghazy, N. A. El-Fishawy, M. M. Hadhoud, M. I. Dessouky, and F. E. A. E.-S. Samie, "An efficient block-by-block svd-based image watermarking scheme," *National Radio Science Conference*, pp. 1–9, 2007.

[21] D. Taskovski, S. Bogdanova, and M. Bogdanov, "Digital watermarking in wavelet domain," *FIRST IEEE BALKAN CONFERENCE ON SIGNAL PROCESSING, COMMUNICATIONS, CIRCUITS, AND SYSTEMS*, 2000.

[22] G. Hai-ying, L. Guo-qiang, L. Xu, and X. Yin, "A robust watermark algorithm for jpeg2000 images," *Fifth International Conference on Information Assurance and Security*, 2009.

[23] D. R. Sans, "Identificao de propriedade em imagens com marcas d'gua no domnio da transformada wavelet," Master's thesis, Universidade Federal do Paran - UFPR, 2008, in portuguese.

[24] E. Ganic and A. M. Eskicioglu, "Robust dwt-svd domain image watermarking: Embedding data in all frequencies," *Proceedings of the 2004 workshop on Multimedia and security*, pp. 166 – 174, 2004.

[25] P. Bao and X. Ma, "Image adaptive watermarking using wavelet domain singular value decomposition," *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, vol. 15, pp. 96–102, 2005.

[26] K. Loukhaoukha and J. Y. Chouinard, "Hybrid watermarking algorithm based on svd and lifting wavelet transform for ownership verification," *11th Canadian Workshop on Information Theory*, pp. 177–182, 2009.

[27] F. B. C. Mendes, "Uma proposta de assinatura digital para imagens por meio de marca d'gua," Master's thesis, Universidade Federal de Uberlandia, 2008, in portuguese.

[28] F. Hartung, J. K. Su, , and B. Girod, "Spread spectrum watermarking: Malicious attacks and counterattacks," pp. 147–158, 1999.

 ISBN:978-1-61208-113-7

Security Analysis of LTE Access Network

Cristina-Elena Vintilă, Victor-Valeriu Patriciu, Ion Bica
Computer Science Department
Military Technical Academy - Bucharest, ROMANIA
cristina.vintila@gmail.com, vip@mta.ro, ibica@mta.ro

Abstract: **The latest technological development in mobile telecommunications is the 4G architecture. Developed and standardized by the 3GPP, this technology proposes significant throughput and security enhancements in comparison with its predecessor, 3G – a 3GPP architecture, as well as with the non-3GPP solutions like WiMAX. Besides the enhancements mentioned above, 4G is a simplified network architecture, flat-IP topology and services-oriented. One of the major simplifications is the base-station architecture, called eNodeB in the 4G terminology, which eliminates the need for a radio resource controller and assumes signaling, control-plane and security functions. It is the mobile device connection to the network and the proxy of all its traffic. This is why the access network is one of the most important areas for network design and optimization and also for security in term of access control, authentication, authorization and accounting. This paper reviews the access network components, the eNodeB security requirements, as defined by 3GPP and analyzes two secure access mechanisms to a 4G network, one via eNB (3GPP access type) and the other one via AP (non-3GPP access type). It also proposes an improvement to the AKA protocol in order to obtain better security.**

Keywords-SAE; LTE; EPC; security; eNodeB; shared-secret; HSS; Diameter; EAP; AKA; J-PAKE

I. INTRODUCTION

The most important and influential telecommunications organizations around the Globe are part of the 3GPP society [20]. The latest technological design for mobile telecommunications that appears to be the future communications architectural baseline is the 4G architecture, also called SAE (System Architecture Evolution). SAE comprises the radio access network, usually referred to as LTE (Long Term Evolution) and the EPC (Evolved Packet Core) the core network of this design, a flap-IP network, highly optimized and secure network, oriented on services.

The figure below describes one of the most common architectural design views, a non-roaming architecture with a 4G mobile device and only 4G access to the network. The entities that appear in this case are the UE (User Equipment), the eNodeB (the antenna), the MME (Mobility Management Entity), the SGW (Serving Gateway), PGW (PDN (Packet Data Network) Gateway), HSS (Home Subscriber Server), PCRF (Policy Charging Rules Function) and a 3G access network where the UE can roam to. This also represents the naming of the interfaces that connect these entities.

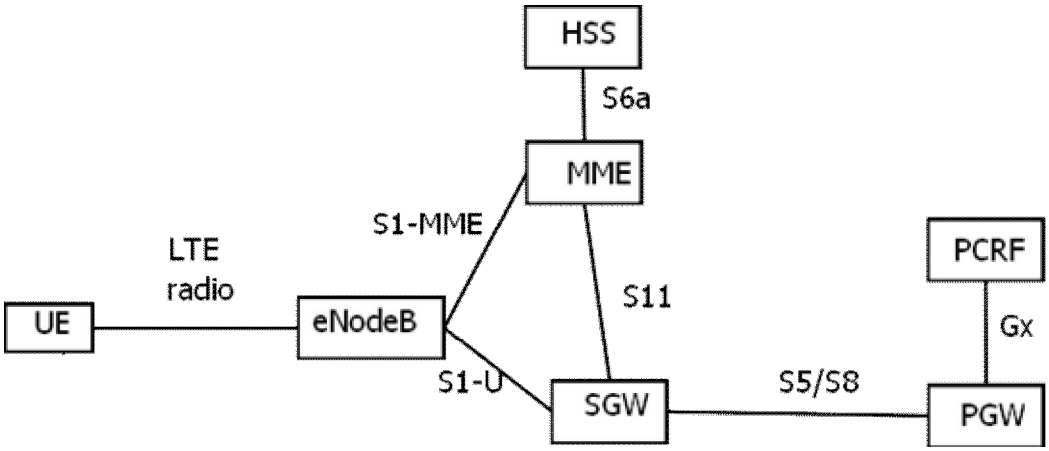


Figure 1. Basic 4G Core Network Architecture

The UE connects to the 4G network by signaling its presence in the eNodeB cell. The cell selection prerequisites are described in [3], [4] and [5]. The process by which an UE chooses a certain antenna (eNodeB) to connect to is called *camping* – the UE *camps* on a cell. The best situation for an UE is to find a so-called *suitable cell* to camp on. This suitable cell is a cell that meets the following requirements: it is part of the selected PLMN (Public Land Mobile Network), part of a registered PLMN or part of the Equivalent PLMN list as per the most recent update from the NAS. Also, the cell should not be barred, not reserved and should not be in the list of forbidden areas for roaming. Once these criteria are met, the UE sends an *Attach Request* message to this eNodeB, asking to attach to the network. This message flows over the LTE

 ISBN:978-1-61208-113-7

radio interface, reaches the eNodeB, and then the eNB sends this message to the MME via the S1-MME interface. The MME verifies the validity of the UE request against the HSS credentials, then selects an appropriate SGW that has access to the PLMN requested by the UE. The PLMN where the UE connects is identified by a string called APN (Access Point Name) preconfigured on the USIM. Once the access request reaches the PGW – which is the UE's anchor point to the desired PLMN, the request is replied with an IP address suitable for this connection and, even more, after interrogating the PCRF about this UE, the PGW may create dedicated bearers for this UE immediately after attach. The decisions on the way are detailed in Section 4.3.8 of [1], which describes the selection functions of each of the core-network entities. Briefly, the HSS drives the selection of the SGW and the SGW on its turn, selects the PGW, based mostly on information already decided upon by the HSS. The MME is selected based on the network topology, the eNB trying to select the MME that minimizes the probability of doing handovers and that provides load balancing with other MMEs.

The Initial Attach process starts with the *Attach Request* message sent by the UE to the eNB selected. This message contains, among other parameters, the IMSI (International Mobile Subscriber Identity) or the old GUTI (Global Unique Temporary Identity), the last TAI (Tracking Area Identifier) if available, PDN Type, PCO (Protocol Configuration Options), Ciphered Options Transfer Flag, KSI-ASME (Key Set Identifier - Access Security Management Entity), NAS (Network Access Server) Sequence Number, NAS-MAC, additional GUTI and P-TMSI (Packet - Temporary Mobile Subscriber Identity) signature.

The PCO means that the UE wants to send some customized information to the network (the PGW may not be in the visited network also), indicating for instance that the UE prefers to obtain the IP address after the default bearer has been activated. If the UE intends to send authentication credentials in the PCO, it must set the Ciphered Options Flag and only send PCO after the authentication and the NAS security have been set up completely. From now on, it is the responsibility of the eNB to proxy the UE's message to the MME. And, once the UE is authenticated, the eNB is also the one responsible of establishing security connections with the UE and the core network in order to protects the UE's traffic at the radio/ethernet border. Being at the border between these two topologies, the eNB is exposed to the security issues arising from both the radio and the IP networks.

II. SECURITY ARCHITECTURE

The 4G architecture defines, in [6], the five main areas concerned with the Security of this design. The first is called Network Access Security and it refers mostly to the radio attacks. The second one is the Network Domain Security and it defines the requirements and rules to prevent attacks over the wire, when exchanging control-plane and user-plane. The third is User Domain Security, dealing with securing the access to mobile terminals, the fourth is the Application Domain Security – which standardizes the set of rules for secure message exchange between applications on clients and servers. The fifth domain defined by this standard is the Visibility and Configurability of Security – set of features that informs the user about a particular security feature and whether this feature is applicable or not to the services this user is trying to access.

The standard [6] describes in Section 5.3 the security requirements necessary for a secure eNB operation environment, as well as for secure eNB functioning. It nevertheless leaves these specifications at a requirements level, permitting the operator to implement the exact protocols he considers for his network; these protocols are compliant to the standard as long as they meet the security requirements defined here, in Sections 11 and 12. The principles are that the eNB should have a mean of securing the cryptographic keys and information inside the device, it should have secure communication links both over the air with the UE and with the MME (via the S1-MME interface), SGW (via the S1-U interface) and other eNBs (via the X2 interfaces, if they exist) for control-plane and user-plane traffic. Also, if the operator has a securely contained environment where these communications happen, he may not implement any precise security measure for the requirements defined here.

The access to a 4G network is done in many ways, most importantly driven by the type of radio medium in place. The most usual procedure is the AKA (Authentication and Key Agreement) Procedure. This happens when the access medium is LTE. The AKA procedure is described in the figure below.

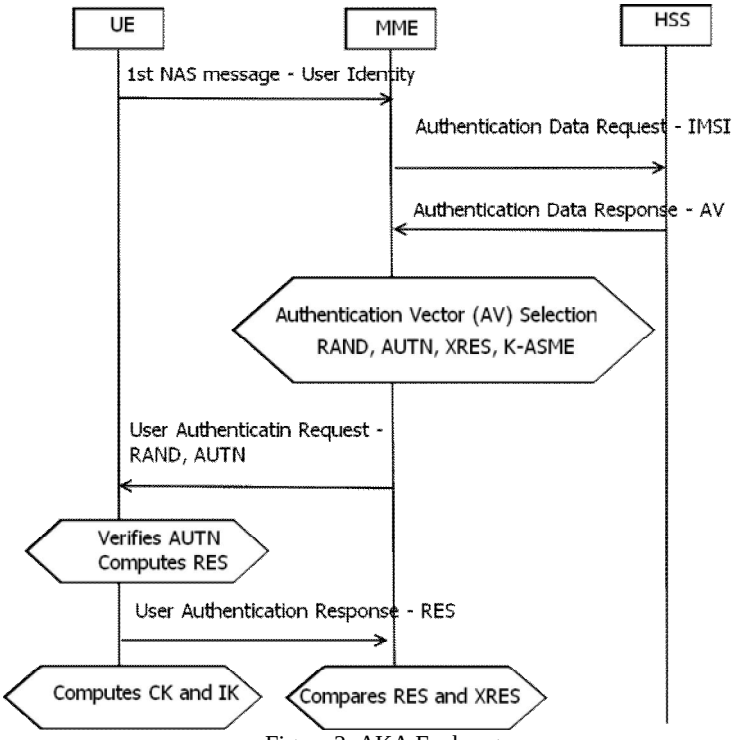


Figure 2. AKA Exchange

The MME here is the authenticator, while the HSS is the authentication server. The communication between the MME and the HSS takes place over S6a and it uses Diameter as a

protocol. The UE (the mobile terminal) has a UICC (Universal Integrated Circuit Card) inside. This circuit stores the K, a shared key located as well on the AuC (Authentication Center) entity part of the HSS. The authentication in 4G is not natively implemented over a PKI infrastructure; it uses the shared-secret symmetric authentication inherited from the 3G UMTS systems, with some improvements.

The purpose of the AKA mechanism is to create keying material for the RRC (Radio Resource Control) signaling, NAS (Non-Access Stratum) signaling and for the user-plane, both ciphering and integrity keys. The first NAS message may be an Attach Request, a Service Request or PDN Connectivity Request message. This message reaches the MME, which should verify the UE's identity. If the UE is new to this network entirely, then the MME asks the UE for its permanent identity – the IMSI. This is considered a security flaw and it is not yet addressed. But, if the UE is not new to this network, but rather reached this MME by means of a TAU (Tracking Area Update) procedure, then this MME should have a GUTI in the message received from the UE. This MME then sends the GUTI and the full TAU message to the previous (old) MME, and this one replies with the actual permanent UE identity – the IMSI and the authentication data for it. The message exchange between the two MME entities takes place over the S10 interface. Also, if the UE roamed to this MME from a 3G network, the current MME tries to connect to the previous management entity of this UE, the (old) SGSN (Serving GPRS Gateway) via the S3 interface and get the IMSI information from there. Otherwise, it tries to derive it and then connects to the HSS via the S6a interface and verifies that the IMSI this UE utilizes is actually valid for this network and may have permission to attach. This message is a Diameter message described in [9].

The HSS gets the AV (Authentication Vector) set and sends it to the MME. This EPS-AV consists of RAND, AUTN, XRES and K-ASME and the HSS, entity that is also called UE's HE (Home Environment) may send multiple sets of AVs to the MME currently serving the UE. The standard recommends that the HSS sends only one set of AVs, but in case it still sends multiple sets, there should be a priority list which the MME should use.

A major improvement when comparing EPS-AKA to UMTS-AKA is that the CK and IK keys never actually have to leave the HSS. The UE signals in its initial message the type of access network he used. If this is E-UTRAN, then a flag called AMF is set to value 1, and this instructs the MME and ultimately the HSS to only send the K-ASME key in the AV reply (along with RAND, AUTN and XRES), but not the CK and IK as well. Also, this K-ASME can be stored in the MME, so, when re-synchronizing the UE's status, the full AKA process may not even have to take place. The MME sends the RAND and AUTN to the UE, then it waits for the response. Here, the eNodeB just forwards these messages back and forth, not participating effectively in the cryptographic exchange. Unlike the GSM, where only the network was authenticating the UE, but not viceversa, the EPS-AKA provides mutual authentication between the UE and the network. Upon receipt of the message, the UE can verify, based on the AUTN, the validity of the reply, computes the RES' and sends this message to the MME. The MME verifies whether XRES equals the RES' and if they are the same, the UE is authenticated. As described, the CK and IK are computed by UICC and HSS independently, they are never sent over the wire in EPS. Also, the HSS sends initial keys to MME and eNB, which are then used by these entities to derive actual keys for NAS, user-plane and RRC traffic.

The figure below depicts two flows: the first one represents the sending of the IMSI in clear-text over the network (the case of the first attach of this UE to the network or when the new MME cannot locate the previous MME) and the second one represents the message exchange between the two MMEs.

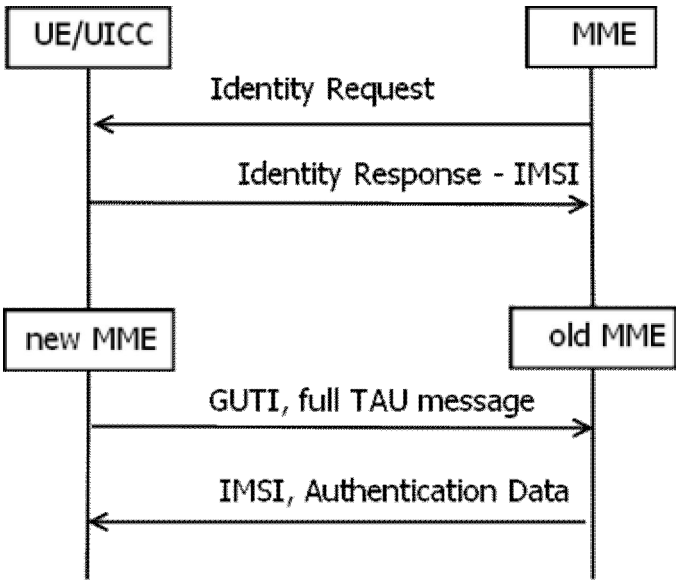


Figure 3. Message exchanges to locate the IMSI

The key hierarchy in 4G is more complicated than in 3G, but it assures this way the protection of the master keys and it also reduces the need for periodic updates, re-generation and transmission of the master keys. There are also special cases for TAU and handover types and also for re-keying that may require special attention, and they are described in [6].

For the non-3GPP access types, the 4G architecture no longer uses the AKA mechanism, but a variation of the EAP with AKA: EAP-AKA (Extensible Authentication Protocol) mechanism. This assumes the presence of a EAP capable phone, the Access Point, which connects to the AAA (Authentication, Authorization, Accounting) server via the Wa interface and the AAA server connects to the HSS via the Wx interface. The EAP-AKA message flow is represented in the picture below. The EAP is a Request/Response type of protocol. When the AP detects the presence of the USIM, it sends this mobile an EAP Identity Request message. The USIM sends back its NAI (Network Address Identifier), which is similar to an e-mail address – RFC 822. The AP forwards this NAI to the 3GPP AAA Server based on the domain name which is part of the NAI – this happens over the Wa interface.

 ISBN:978-1-61208-113-7

Then the AAA server verifies whether it has a valid and unused AV for this USIM. If so, it sends this AV and AKA Challenge for the USIM, back to the AP. If there is no available AV for this USIM, the AAA server contacts the HSS server via the Wx interface, retrieves the AV and then continues to the AKA Challenge. The rest of the process is similar to the EPS-AKA, with the only observation that it takes place over the EAP framework.

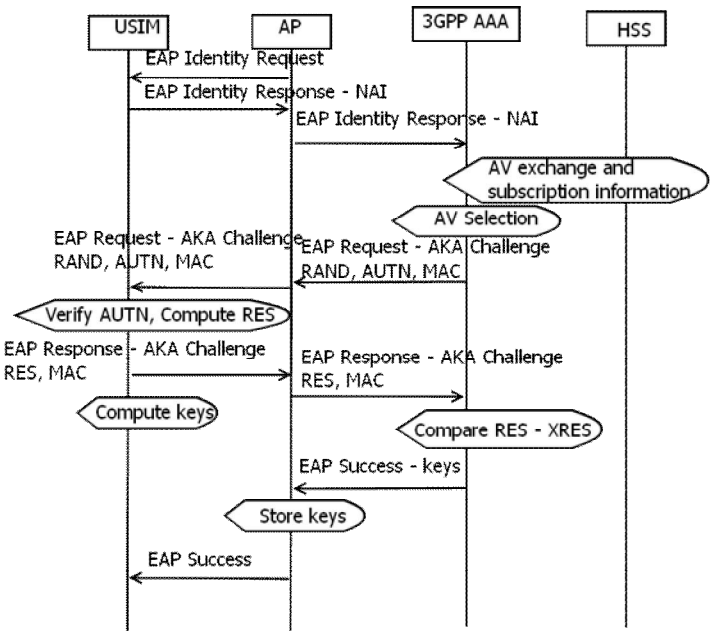


Figure 4. EPS EAP-AKA exchange

When comparing pre-4G authentication methods, there are several aspects that may be observed. One of them is the authentication method. This is very similar for 3G (UMTS), 3.5 G (HSDPA) and 3.75G (HSDPA+), as all of them use the AKA mechanism. The differences appear in the actual implementation: the 3G implementation specifies that the CK and IK keys from the AuC (Authentication Center) part of the HE (Home Environment) are actually being sent to the SGSN at the moment the SGSN downloads the Authentication Vectors from this database. This never happens in 4G, where the key hierarchy is more complicated and the only key downloaded from the HSS to the MME is the K-ASME key. This is also a security improvement in 4G in comparison to 3G, because the CK and IK keys should not leave the AuC, but only be derived independently by the UICC and HSS. In both 3G and 4G security architectures, there are multiple AVs (Authentication Vectors) available in the authentication part of the subscribers database. All these authentication vectors may be downloaded initially by the authentication entity (SGSN, MME respectively), a certain AV being used for a single round of authentication. The order in which these AVs are used is determined in both architectures by a sequence number. The CK||IK pair is derived by the UICC in 3G, and the SGSN only selects this pair from the authentication data received from the AuC. In 4G, the MME receives only its K-ASME from the HSS, and it then derives, together with the UE, the K-NASint and K-NASenc from the K-ASME. The following figure describes the key hierarchy in the 4G architecture.

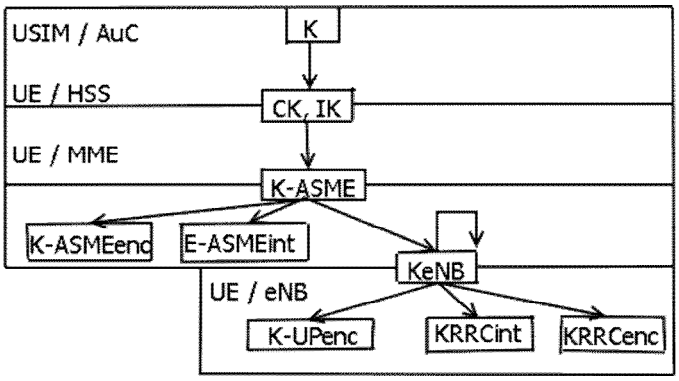


Figure 5. EPS key hierarchy [6]

The keys derived in the classical 4G AKA procedures are the following: K-NASenc (encryption key for NAS traffic), K-NASint (integrity key for NAS traffic), KeNB (derived by MME and eNB), KUPenc (encryption key for the user-plane data, derived by MME and eNB from KeNB), KRRCint and KRRCenc (integrity check, respectively encryption key derived by MME and eNB from KeNB, used for securing the Radio Resource Control traffic).

The sections in this figure describe which entities are involved in a particular key generation process; there are 6 keys derived from the EPS authentication mechanism. This key generation feature introduced in 4G improves the speed of the re-authentication procedures and also the refreshing process of the keys. As the K-ASME may be used a master key in further service requests authentication, the MME no longer has to download the authentication data from the HSS and it can also avoid re-synchronization issues. HSDPA and HSDPA+ follow the 3G procedures and mechanisms.

The entire authentication information that is stored in the UICC and its corresponding associates from the network side is called “security context”; a security context consists of the NAS (Non Access Stratum) and AS (Access Stratum) security contexts. The AS security context has the cryptographic keys and chaining information for the next hop access key derivation, but the entire AS context exist only when the radio bearers established are cryptographically protected. The NAS context consists of the K-ASME with the associated key set identifier, the UE security capabilities and the uplink and downlink NAS count values, used for each EPS security context. The 3G architecture also has the concept of security contexts, and this becomes very important when 3G and 4G devices and network entities interoperate. A 3G device that was initially attached to a 3G network has received a set of security content that is stored in the UICC. This information is considered a partial and legacy security context in the 4G environment. For a security context to be considered full, the MME should have the K-NASenc and K-NASint keys, which are obtained when the 3G device handovers to the 4G network. In this handover case, the legacy security context is

 ISBN:978-1-61208-113-7

referred to as mapped security context. A NAS security context of a mapped security context is always full and it is current (which means it is the most recently activated context). A summary of the types of security contexts that exist in the 3G – 4G interoperation is in the following table.

TABLE I. TYPES OF SECURITY CONTEXTS

AGE/EFFECT	CURRENT	NON-CURRENT
FULL	NATIVE / MAPPED	NATIVE
PARTIAL	X	NATIVE

A native security context is the one established at the EPS-AKA procedure successful completion.

III. SECURITY ISSUES AND ASSESSMENT

The security analysis in a mobile network expands from the radio access network until the core network and services. The most common threats are related to the following security prerogatives:

- authentication: the network must be sure that the person accessing a certain service is the one pretending to be and paying for this service
- confidentiality of data: the user and the network must be sure nobody un-authorized is viewing or accessing the user's data
- confidentiality of location: the user's location must not be known by anybody un-authorized
- denial of service: the user and the network must be sure that nobody interferes with a user's session, nor high-jacks it
- impersonation: the network and the user must be sure that no other user is pretending to be the actual registered user, nor this user can access services available to the actual registered user

The EPS security mechanisms should be able to enforce the above principles, and everything starts with the access level. The NDS (Network Domain Security) enforces these principles as well, but at a different level. When talking about User Domain Security, the Access Level is the first line of defense against attacks. For this, the EPS provides mutual authentication via the EPS-AKA mechanism or EPS-EAP-AKA (for non-3GPP access types). Using the keys generated after the AKA mechanism, all the RRC, NAS and user-plane signaling and data-planes are protected by secure encapsulation and integrity protected. Even though a flaw of the EPS-AKA: the transmission of the permanent identity IMSI over the air at Initial Attach exchange, EPS implements the GUTI value that is sent over the air instead of the IMSI, so that the AKA mechanism also provides identity protection. The only cases that require the transmission of the IMSI are the first Initial Attach and the attach after the core network entities are de-synchronized. This vulnerability opens the door for a man-in-the-middle attack that can take place once the IMSI is captured. Another vulnerability identified within the EAP-AKA, but that manifests in consequence with EPS-AKA also is the lack of PFS support. The PFS – Perfect Forward Secrecy is an attribute of the mechanism's that assures the secrecy of a set of session keys even if a previous set of keys has been compromised.

A solution for the PFS vulnerability can be the use of an algorithm that has the PFS built-in. This algorithm is the Diffie-Hellman key exchange. Still, DH does not provide mutual authentication by itself. Another algorithm, based on the J-PAKE mechanism, can be used. The J-PAKE mechanism has the following properties – as described in [19]:

- off-line dictionary attack resistance – it does not leak any password verification information to a passive attacker
- forward secrecy – it produces session keys that remain secure even when the password is later disclosed
- known-key security – it prevents a disclosed session key from affecting the security of other sessions
- on-line dictionary attach resistance – it limits an active attacker to test only one password per protocol execution

One solution for using the Juggling scheme in the UE's authentication to the network is to replace part of the AKA protocol with the J-PAKE protocol.

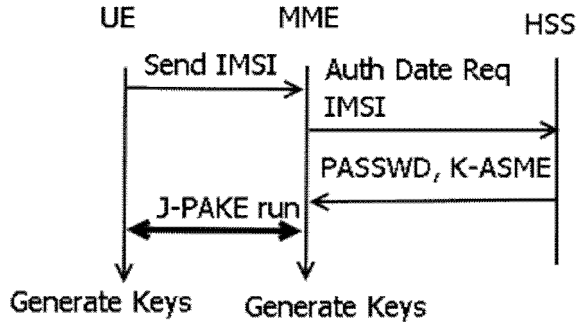


Figure 6. Simplified J-PAKE usage in 4G authentication

This solution does not cover the first security issue present in AKA, the identity protection. This aspect will be debated in a separate article. This solution also assumes the existence of a secure communication channel between the MME and the HSS. Once the IMSI is sent to the HSS by the MME, the HSS will return the shared key it has for this UE, secured via the S6a interface. Having the password, the MME will run the J-PAKE protocol with the UE, the eNB serving as simple relay agent; the UE proves, via the J-PAKE rounds, its knowledge of the password, and at this moment, the UE is considered authenticated by the network and it has also authenticated the network, This is a mutual authentication, resistant to the 4 security aspects listed above. It improves on the AKA algorithm by providing forward secrecy for the keys resulted from this negotiation. The generation of the EPS key hierarchy is not impacted by the authentication mechanism, the generation remains the same as described by the standard.

A more detailed comparison and simulation on the effectiveness of the J-PAKE method versus the AKA method is in progress.

J-PAKE is a password authentication keying agreement protocol, a method to provide zero-knowledge proof using a shared key that is never sent over the transmission medium. One of the first algorithms of this type is the EKE (Encryption Key Exchange) protocol, which has been proved to have many

 ISBN:978-1-61208-113-7

flaws. SPEKE (Simple Password Exponential Key Exchange) is one of the protocols that improved the EKE variant. SPEKE, along with EKE and J-PAKE are all balanced versions of password-authenticated key agreement. This variant uses the same password to authenticate both peers. There is another variant of PAKE, called augmented PAKE. This variant is usable in a client/server environment. Here, a brute-force attack on the connection is more inconvenient and the representative protocols for this variant are B-SPEKE and SRP (Secure Remote Password protocol). J-PAKE improves on the limitations of S-PEKE, limitations that are already observed in the BlackBerry implementation. The actual protocol – level security comparison between J-PAKE and SPEKE are described in the J-PAKE presentation paper. The conclusions are that EKE does not fulfill the off-line dictionary attack resistance requirement, while the SPEKE does not fulfill the on-line dictionary attack resistance requirement.

IV. Conclusions and Future work

This paper presented the basic 4G architecture and reviewed the main security domains used to classify and integrate the security aspects to this design. The paper focused on the access-level security issues that may arise at the eNodeB, UE and MME level. As the connection to the network poses the most issues when talking about User Domain security, this paper analyzed the authentication process of the UE connecting to the 4G network. There are two most common cases when talking about initial attach, the attach of a plain 4G device and the attach of non-3GPP device.

The paper identified two issues that appear almost every time: the lack of identity protection at the first initial attach and the lack of perfect forward secrecy for the AKA mechanism, inherited also in the EAP-AKA mechanism specific to the authentication of the non-3GPP devices. At least for the perfect forward secrecy issue, we have proposed the usage of the J-PAKE protocol in the authentication process, instead of the AKA protocol, which we consider a flexible and lightweight mechanism, suited for use in the mobile device environment.

The future articles will describe the comparison and a possible simulation of the efficiency of J-PAKE in comparison with SPEKE and other similar balanced PEKE algorithms, as well as measure the efficiency of this protocol as the main authentication algorithm in the 4G authentication process.

Further on, this study continues with the analysis of the IMSI identity protection mechanism and proposes a solution for the complete identity protection even for the first initial attach process.

References

[1] TS 23.401, GPRS Enhancements for E-UTRAN access, http://www.3gpp.org/ftp/Specs/archive/23_series/23.401/23401-a10.zip [retrieved: November 2010]
[2] TS 23.122, NAS Functions related to Mobile Stations in idle mode, http://www.3gpp.org/ftp/Specs/archive/23_series/23.122/23122-a10.zip [retrieved: November 2010]
[3] TS 36.300, E-UTRAN Overall Description, http://www.3gpp.org/ftp/Specs/archive/36_series/36.300/36300-a10.zip [retrieved: November 2010]
[4] TS 43.022, Functions of the MS in idle mode and group receive mode, http://www.3gpp.org/ftp/Specs/archive/43_series/43.022/43022-920.zip [retrieved: November 2010]
[5] TS 25.304, UE Procedures in idle mode and procedures for cell reselection in connected mode, http://www.3gpp.org/ftp/Specs/archive/25_series/25.304/25304-930.zip [retrieved: November 2010]
[6] TS 33.401, SAE - Security Architecture, http://www.3gpp.org/ftp/Specs/archive/33_series/33.401/33401-950.zip [retrieved: November 2010]
[7] TS 33.310, Network Domain Security; Authentication Framework, http://www.3gpp.org/ftp/Specs/archive/33_series/33.310/33310-a10.zip [retrieved: November 2010]
[8] TS 33.102, 3G Security Architecture, http://www.3gpp.org/ftp/Specs/archive/33_series/33.102/33102-930.zip [retrieved: November 2010]
[9] RFC 5516, Diameter Command Code Registration for the Third Generation Partnership Project (3GPP) Evolved Packet System (EPS), http://tools.ietf.org/html/rfc5516 , April 2009 [retrieved: November 2010]
[10] TS 29.272, MME related interfaces based on Diameter, http://www.3gpp.org/ftp/Specs/archive/29_series/29.272/29272-a00.zip [retrieved: November 2010]
[11] Tech-Invite, http://tech-invite.com/, [retrieved: March 2010]
[12] TS 29.294, Tunneling Protocol for Control plane (GTPv2-C), http://www.3gpp.org/ftp/Specs/archive/29_series/29.274/29274-a00.zip [retrieved: November 2010]
[13] TS 33.220, Generic Authentication Architecture; Generic Bootstrapping Authentication, http://www.3gpp.org/ftp/Specs/archive/33_series/33.220/33220-a00.zip [retrieved: November 2010]
[14] TR 33.919, Generic Authentication Architecture – System Overview, http://www.3gpp.org/ftp/Specs/archive/33_series/33.919/33919-910.zip [retrieved: November 2010]
[15] TS 33.221, Support for Subscriber Certificates, http://www.3gpp.org/ftp/Specs/archive/33_series/33.221/33221-910.zip [retrieved: November 2010]
[16] Han-Cheng Hsiang and Weu-Kuan Shih, “Efficient Remote Mutual Authentication and Key Agreement with Perfect Forward Secrecy”, Information Technology Journal 8, 2009, Asian Network for Scientific Information [retrieved: November 2010]
[17] RFC 4187, EAP Method for 3GPP AKA, http://tools.ietf.org/html/rfc4187 [retrieved: November 2010]
[18] RFC 2631, Diffie-Hellman Key Agreement Method, http://tools.ietf.org/html/rfc2631 [retrieved: November 2010]
[19] F. Hao and P.Ryan, “Password Authenticated Key Exchange by Juggling”, Proceedings of the 16th International Workshop on Security Protocols,2008, http://grouper.ieee.org/groups/1363/Research/contributions/hao-ryan-2008.pdf [retrieved: November 2010]
[20]3GPP, http://3gpp.org/-partners-, [retrieved: November 2010]
[21] Georgios Kambourakis, Angelos Rouskas, and Stefanos Gritzalis, "Performance Evaluation of Public Key-Based Authentication in Future Mobile Communication Systems", August 2004, EURASIP Journal on Wireless Communications and Networking [retrieved: November 2010]
[22] Qiang Tang and Chris J. Mitchell, “On the security of some password-based key agreement schemes”, 23rd May 2005 [retrieved: November 2010]

 ISBN:978-1-61208-113-7

Anonymous Key Issuing Protocol for Distributed Sakai-Kasahara Identity-based Scheme

Amar SIAD
Laboratoire Analyse Géométrie et applications LAGA-Paris 13
Université Paris 8, 2 rue de la Liberté 93526
SAINT-DENIS, France
siad@math.univ-paris13.fr

Moncef AMARA
Université Paris 8, 2 rue de la Liberté 93526
SAINT-DENIS, France
amara02@etud.univ-paris8.fr

***Abstract*—Practical implementations of identity based cryptosystems are faced to key escrow problem, which is not always a good property in many realistic scenarios. Thus, efficient key issuing protocols are needed to generate and deliver user's private keys in secure manner without leakage. Three major approaches exist in the literature, we are interested in two of them. The first one suggested the distribution of the master secret key over multiple authorities. The second approach, concerned about user privacy, and proposed to generate and deliver user's private keys in an anonymous manner. Each one of the above approaches has its own drawbacks and key escrow problem is steal an issue in identity-based systems. In this paper, we design a new framework that combines the two approaches above to solve key escrow problem and single point of failure in Identity-based Encryption systems by allowing privacy-preserving propriety. As instantiation, we construct an anonymous key issuing protocol for the distributed sakai-kasahara IBE scheme presented recently by *Kate and Goldberg* based on the anonymous key issuing protocol proposed by *Chow*, along with a security analysis.**

Keywords-Key issuing protocols; Distributed key generation; Anonymous IBE.

I. INTRODUCTION

Traditional Public Key Infrastructure PKI, supporting Public Key Cryptography PKC, provided mechanisms required for certificate issuing, maintain, and revocation. Thus, it has succeeded in many applications by managing the trust between different entities. However, PKI is not a perfect solution, and many problems steal subsists due to the administrative burden of certificates, revocation lists or trees, and cross-domain certification.

In 1984, Shamir [1] proposed a novel concept called Identity Based Public Key Cryptography (ID-PKC), where the original motivation is to simplify certificate management in PKI-based systems. The main idea of so called ID-PKC is to derive users public key from his identity information whereas the private key is generated by a third party called Private Key Generator (PKG) and issued to the user via a secure channel. Shamir presented an identity based signature system (IBS) using RSA and conjectured that encryption systems could be constructed. Compared to traditional Public Key Cryptography, ID-PKC present the advantage of simplified key distribution and management (no need for certificates). However, it suffers from an inherent drawback of key escrow, where the PKG could decrypt any message addressed to a user by generating that user's private key. Moreover, it requires a secure channel for users' private keys issuance.

Since the shamir's challenge, the cryptographic community had to wait until the turn of the century to see practical constructions of ID-PKC systems, considered thus far an open problem. The first scheme by Cocks [2] using the quadratic residues, whereas the second one by Boneh and Franklin [3] using Weil pairings on elliptic curves.

Boneh and Franklin construction have widely opened doors to an important development in recent years. Thus, a flurry of schemes have been proposed, improved, proven secure, and security formal models have been more and more strengthen. However, the deployment of practical ID-based systems have not followed the same rhythm of these theoretical improvements and the few systems' implementations proposed deal with a set of particular limited scenarios. [4] pointed-out that the deployment of an ID-based system requires an infrastructure as complex as a PKI. *Chen et al.* [5] presented a hybrid scheme combining traditional PKI with ID-PKC in a multi-authority environment.

Interoperability issues of ID-PKC and PKI are also discussed in [6]. Whereas many works studied key issuing protocols [3], [7], [8], [9], [10], [11] presenting wide rage solutions of key escrow problem, but none of the proposed solution is perfect and key escrow is steal an issue facing the deployment of ID-based systems. In the same scope [12] developed an architecture model for distributed PKG, using PKI, for internet applications. Recently, Chow [10] exploited the anonymity propriety to fight against key escrow problem by defining an anonymous key issuing protocol for Gentry scheme.

We organize the rest of the paper as follows. In Section II, we give related work and our contribution. In Section III, we give some preliminaries. In Section IV, we define the general framework and architecture along with security requirement. In Section V, we present a construction of a

 ISBN:978-1-61208-113-7

distributed AKI for SK-IBE scheme. Finally, we conclude in Section VI.

II. Related work

Key escrow problem made the deployment of practical ID-PKC cryptosystems limited to small and relatively closed organisations where the trust in PKG is very heigh. To tackle this restriction and extend the use of ID-PKC in scenarios equivalent to the ones of PKI-based PKC, key issuing protocols are studied. These protocols allow the user to have his key without leakage. We classify key issuing protocols into three main categories described hereafter.

Multi-authorities and distributed protocols. in addition of key escrow problem, this approach deal with the problem of single point of failure. [3], [13], [14], [15], [16] proposed different but related approaches to split the master secret key to multiple $\mathcal{KGC}$s. The user obtains a partial private key from each $\mathcal{KGC}$ and reconstructs his private key in threshold manner. [7], [17] used the concept of key privacy authorities $\mathcal{KPA}$ to deliver user's private key in blinded manner using a single $\mathcal{KGC}$ and multiple $\mathcal{KPA}$. Recently, *Geisler and Smart* [11] proposed distribution version of sakai-kasahara based systems, *Kate and Goldberg* [18] developed a distributed private-key generators for three IBE schemes along with their security proofs.

Anonymous protocols. Anonymous key issuing protocols where first considered by *Sui et al.* [19] where they separate authentication phase from key issuing by using a database to store identities and corresponding passwords, whereas the fact of using the database gives the $\mathcal{KGC}$ the capability to link key requests with user's identity and break the anonymity of the proposed protocol. Recently, chow [10] extended the anonymity notion to fight against adversaries who hold the master key and proposed an anonymous key issuing protocol for a modified gentry IBE scheme. However, as chow pointed out, the proposed protocol has a major drawback where the $\mathcal{KGC}$ can generate all possible user private keys by guessing user identities according to some dictionary.

User-chosen secret information. Another approach have been introduced by [8], [9] who proposed respectively the concept of Certificate-Based Encryption (CB-PKC) and Certificateless Cryptography (CL-PKC). These two solutions avoid successfully the escrow problem by combining advantages of traditional PKC and ID-PKC to create a hybrid model. However, as already evoked in [7] these approach loose the main propriety of an IBE system in which the user public key is derived from his identity, and thus are not considered purely IBE systems.

A. Our contribution.

The main idea behind our protocol is to combine multi-authorities approaches with anonymous protocols in order to develop a new class of protocols. In [10], the use of an anonymous protocol prevent the $\mathcal{KGC}$, or an adversary having access to the master key, from linking user's identity to the private key generated for that identity. However, the $\mathcal{KGC}$ can still generate private keys for identities of his choice (ie. by guessing users' identities) and then proceed by an off-line analysis of messages flow by trying to decrypt messages using the key generated. To overcome this drawback, we propose to extend the anonymous key issuing protocol in [10] to prevent the $\mathcal{KGC}$ from this capability by distributing the $\mathcal{KGC}$ master secret key over multiple authorities in conjunction with a certification authority CA to authenticate users and deliver new kind of certificates by signing on a committed value of the user identity, the same way as in [10]. The certificate will be presented by user to each one of the n $\mathcal{KGC}$ to get his private key. This new architecture, will solve key escrow problem and single point of failure. Our contribution can also seen as an extension of the distributed protocols in [11], [18] to support user anonymity.

III. Preliminaries

A. Bilinear pairings

Let $\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T$ be cyclic groups of prime order q, P_1 a generator of $\mathbb{G}_1$, and P_2 a generator of $\mathbb{G}_2$. A bilinear pairing e is a map defined by e: $\mathbb{G}_1$x $\mathbb{G}_2 \rightarrow \mathbb{G}_T$ with the following properties:

1) Bilinear : $e(aP, bQ) = e(P, Q)^{ab}$, $\forall P \in \mathbb{G}_1$, $Q \in \mathbb{G}_2$ and $a, b \in \mathbb{Z}_q^*$.
2) Non degeneracy : $e(P_1, P_2) \neq 1$.
3) Efficiently computable: there exists an efficient algorithm to compute $e(P, Q)$ for all $P \in \mathbb{G}_1$, $Q \in \mathbb{G}_2$.

B. Distributed key generation

Hereafter we give a quick review of the distributed computation primitives used in this work, for more details we refer the reader to [18]. Distributed key generation DKG is introduced by Pederson [20], who developed a DKG that requires no dealer. An (n, t) DKG is composed of n nodes that generate a secret $z \in \mathbb{Z}_p$ in a distributed fashion. Each node gets a share $z_i \in \mathbb{Z}_p$ such that any subset of size greater than t could reconstruct the secret.

Shares Generation. depending on weather we use discrete logarithm (Dlog) or Pedersen (Ped) commitments, nodes use one of the two protocol $(Random_{Dlog}(), Random_{Ped}())$ to generate shares of a secret $z \in \mathbb{Z}_p$ chosen jointly at random.

1) $\left(\mathcal{C}_{\langle g \rangle}^{(z)}, z_i\right) = Random_{DLog}(n, t, g)$
2) $\left(\mathcal{C}_{\langle g,h \rangle}^{(z,z')}, \left[\mathcal{C}_{\langle g \rangle}^{(z)}, NIZKPK_{\equiv com}\right], z_i, z_i'\right) = Random_{Ped}(n, t, g, h)$.

Recall that $\left(\mathcal{C}_{\langle g \rangle}^{(z)}\right) = [g^z, g^{\phi(1)}, ..., g^{\phi(n)}]$ and $\left(\mathcal{C}_{\langle g,h \rangle}^{(z,z')}\right) = [g^z h^{z'}, g^{\phi(1)} h^{\phi'(1)}, ..., g^{\phi(n)} h^{\phi'(n)}]$ are respectively Discret log and Pedersen commitment victors for z, and $\phi, \phi' \in \mathbb{Z}_p[x]$ are polynomials of degree t where $\phi(0) = z, \phi'(0) = z'$, $\phi(i) = z_i$, and $\phi'(i) = z_i'$.

Distributed Multiplication. for distributed multiplication we use the second protocol from [18] that uses a multiplication protocol against computational adversaries with a non-interactive proof of knowledge defined as follows.
$\left(\mathcal{C}_{\langle \hat{g},\hat{h} \rangle}^{(\alpha\beta,\alpha\beta')}, (\alpha\beta)_i, (\alpha\beta')_i\right) = Mul_{Ped}(n, t, \hat{g}, \hat{h}, \left(\mathcal{C}_{\langle g \rangle}^{(\alpha)}, \alpha_i\right), \left(\mathcal{C}_{\langle \hat{g},\hat{h} \rangle}^{(\beta,\beta')}, \beta_i, \beta_i'\right))$.
By this each node computes locally the share of the product of two shared secrets α, β.

C. Sakai-Kasahara-IBE

SK-IBE Setup(λ): Given the security parameter λ, the parameter generator follows the steps.

1) Generate three cyclic groups $\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T$ of prime order q, a bilinear pairing map $e : \mathbb{G}_1$ x $\mathbb{G}_2 \rightarrow \mathbb{G}_T$ and random generators $(g, \hat{g})$ for respectively $\mathbb{G}_1$, $\mathbb{G}_2$.
2) Pick four cryptographic hash functions $H_1 : \{0,1\}^* \longrightarrow \mathbb{Z}_p$, $H_2 : \mathbb{G}_2 \longrightarrow \{0,1\}^n$, $H_3 : \{0,1\}^n$x$\{0,1\}^n \rightarrow \mathbb{Z}_q^*$ and $H_4 : \{0,1\}^n \rightarrow \{0,1\}^n$ for $n > 0$.
3) Pick a random $s \in \mathbb{Z}_q^*$ and compute $pk = g^s$

The public parameters are $params = (q, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, e, n, g, \hat{g}, g^s, H_1, H_2, H_3, H_4)$

SK-IBE Extract(msk, ID): the private key d_{ID} of user having ID as identity is computed bay:$d_{ID} = \hat{g}^{\frac{1}{s+H_1(ID)}}$

SK-IBE Encryption(mpk, ID, m) : to encrypt a k' bit length message M, the sender picks at random $\sigma \in \{0,1\}^{k'}$, computes $r = H_3(\sigma, M)$, $h_{ID} = H_1(ID)$ and sends the cyphertext $C = (u, v, w) = ((g^s g^{h_{ID}})^r, \sigma \oplus H_2(e(g, \hat{g})^r), M \oplus H_4(\sigma))$ to the recipient.

SK-IBE Decryption(d_{ID}, c): to decrypt the cyphertext $C = (u, v, w)$ with the private key d_{id}, the receiver successively computes $\sigma = v \oplus H_2(e(u, d_{id}))$, $M = \sigma \oplus H_4(\sigma)$, $r = H_3(\sigma, M)$. If $(g^s g^{h_{ID}})^r \neq u$ then C is rejected else M is a valid message.

D. Anonymity in IBE

Anonymity against user attacks for IBE was introduced first by *Abdella et al.* [21] similarly to semantic security. The attacker's goal is to distinguish the intended recipient of a ciphertext between two chosen identities.The previous definition of anonymity cannot provide security against $\mathcal{KGC}$ attacks. Recently, this notion was strengthen and extended to handle $\mathcal{KGC}$ attacks by two independent works [22], [10]. Izabachne and pointcheval [22] called it KwrtA-Anonymity (Key Anonymity with respect to the Authority) and applied it in password-authenticated key exchange. Whereas, Chow [10] called it $\mathcal{ACI} - \mathcal{KGC}$ (Anonymous Cyphertext Indistinguishability) and used it to fight against key escrow.

IV. General framework

We extend the framework given in [10] to support the distributed architecture. We assume the existence of a certification authority $\mathcal{CA}$ and multiple key generation centres.

A. Entities and Their Roles

The entities involved the new architecture are as follows.

- $\mathcal{CA}$**:** certification authority is a trusted authority in the standard PKI based model. The CA is responsible for checking users identities and certificate issuing, it is clear that $\mathcal{CA}$ holds the identity list of all users in the system. $\mathcal{CA}$ has a master secret key sk_{cert} and the corresponding public key pk_{cert}.
- n $\mathcal{KGC}$**:** multiple authorities for user key generation using (n, t) threshold secret sharing scheme and without knowing the identity of the user. Each $\mathcal{KGC}$ has a secret key s_i and the corresponding public key pk_i. We make the assumption that $\mathcal{CA}$ doesn't collude with $\mathcal{KGC}s$, otherwise the user anonymity can be broken.
- **User:** he should first present authenticate himself to the $\mathcal{CA}$ which issue a certificate on a commitment on the user identity. Then he presents his certificate to each one of the t $\mathcal{KGC}$s to get partial private keys. Finally, the user compute his private key by interpolation.

B. AKI for distributed IBE

Définition 4.1: **(AKI for (n, t) IBE)** an anonymous key issuing protocol for (n, t) IBE scheme consists of components ($\mathcal{SIGN}, \mathcal{P} - \mathcal{SIGN}, \mathcal{DKG}, \mathcal{C}$) specified as follows:

1) $\mathcal{SIGN}$: signature scheme run by $\mathcal{CA}$ to generate user certificate. The certificate is delivered to the user securely and is presented by user to $\mathcal{KGC}$. Note that the certificate doesn't contain user's name and not used anywhere else in the system.
2) $\mathcal{P} - \mathcal{SIGN}$: p-signature scheme [23] that allows the user, with a private input, to get a signature on a commuted value of the identity without revealing it to the signer. Note that p-signature is a primitive that uses secure two-party computation protocol on committed values.
3) $\mathcal{DKG}$: Distributed key generation protocol, that takes as input the security parameter λ , the threshold parameters (t,n) and outputs for each player P_i (for $i = 1, ..., n$) a share s_i of the master secret key s and

a public-key vector K_{pub} of a master public key and n public-key shares.;

4) $\mathcal{C}$: non interactif commitment scheme.

More formally, An AKI-protocol for (n, t) IBE scheme is defined by the following algorithms as follows:

$(pk_{CA}, sk_{CA}, cert_{CA}) \leftarrow$ **SetupCA(λ):** probabilistic algorithm executed by $\mathcal{CA}$, it takes as input security parameter λ and returns $\mathcal{CA}$ public key pk_{CA}, master secret key sk_{CA}, and $\mathcal{CA}$ certificate $cert_{CA}$.

$(cert_U, open) \leftarrow$ **CertIssue(sk_{CA}, ID):** probabilistic algorithm executed by $\mathcal{CA}$ to deliver certificates to users. It takes as input $\mathcal{CA}$ secret key sk_{CA}, user identity ID and returns user certificate $cert_U = (sig, comm, open)$, where $open$ is chosen at random from the decommitment-string space and sig is a signature on $comm = Commit(H(ID), open)$, where H is a hash function.

$(s_1, pk_1, ..., s_n, pk_n, pk) \leftarrow$ **DKeyGen(λ, t, n):** distributed key generation protocol runs between the n $\mathcal{KGC}$ and results in each $\mathcal{KGC}$ obtaining a share $s_i \in \mathbb{Z}_q$ of the master secret s. The tuple $(pk, pk_1, ..., pk_n)$ is the system public-key.

ObtainKey$(\mathcal{U}(params, id, cert_U, open)) \leftrightarrow$ **IssueKey** $(\mathcal{KGC}_i(params, s_i, cert_U))$ **:** an interactive protocol, using a secure two-party computation protocol, executed between user $\mathcal{U}$ and $\mathcal{KGC}_i$ for $i = 1, ...t+1$ ($t+1$ out of n $\mathcal{KGC}$). $\mathcal{U}$ takes as input master public key mpk, the identity id, certificate $cert$, opening information $open$ and gets a partial secret key $d_{id}^{(i)}$ as output. $\mathcal{KGC}_i$ takes as input master secret key s_i, user certificate $cert_U$ and gets nothing as output.

ReconstructKey($d_{id}^{(1)}, ..., d_{id}^{(t+1)}$): upon receiving $t+1$ partial private key, user reconstructs his private keys in threshold manner.

C. Security requirements

Définition 4.2: **(Secure AKI :)** an anonymous key issuing protocol for (n, t) IBE is secure if: *(1)* p-signature scheme is unforgeable and satisfies signer privacy and user privacy; *(2)* DKG protocol satisfies correctness and secrecy; *(3)* commitment scheme is perfectly binding and strongly computationally hiding.

V. Construction

chow [10] argued that SK-IBE can be made $\mathcal{ACI} - \mathcal{KGC}$ the same way as gentry scheme by separating parameters generation from key generation. Admitting this fact, hereafter we modify SK-IBE to support anonymity against $\mathcal{KGC}$ and give an anonymous key issuing protocol. p-signature scheme from [23] is used in [10] to construct an anonymous key issuing protocol for modified gentry scheme, assuming the same framework architecture, we adapt this protocol to a modified version of SK-IBE scheme that supports $\mathcal{ACI} - \mathcal{KGC}$.

A. AKI for SK-IBE scheme

Setup public parameters are generated by trusted initializer as follows. $params = (q, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, e, n, g, \hat{g}, H_1, H_2, H_3, H_4)$

SK-IBE KeyGen: the $\mathcal{KGC}$ randomly chooses an exponent s and computes g^s. (s, g^s) is the private/public key pair of the $\mathcal{KGC}$.

Encrypt(), Decrypt(): these algorithms remain as in the original version, whereas the private key extraction algorithm **Extract()** is modified as follows.

ObtainKey$(\mathcal{U}(params, id, cert_U, open)) \leftrightarrow$ **IssueKey** $(\mathcal{KGC}(\ params, msk, cert_U))$ **:**

- User presents his certificate $cert_U$ to $\mathcal{KGC}$, the latter verifies certificate signature using the CA public key pk_{cert} if certificate verification fail it aborts the protocol.
- The user chooses at random ρ in $\mathbb{Z}_q$;
- The user and $\mathcal{KGC}$ engage in a secure two-party computational protocol[24], where the user's private input is $(\rho, H_1(ID), open)$, and the $\mathcal{KGC}$'s private input is msk. As result, the $\mathcal{KGC}$ gets a private output which is either $x = (x + H_1(ID))\rho$ if $comm = commit(H_1(ID), open)$ or $\perp$ in this case the $\mathcal{KGC}$ aborts.
- if $x \neq \perp$ the $\mathcal{KGC}$ send $\sigma' = \hat{g}^{\frac{1}{x}}$ to the user;
- the user computes, upon receiving σ', $\sigma = (\sigma')^{\rho} = \hat{g}^{\frac{1}{msk + H_1(ID)}}$.

B. Security analysis

Recall the definition of p-signature, which is a signature on a committed message without revealing the message using a secure two-party computation protocol on committed inputs, the user private key in SK-IBE scheme can be seen as the first p-signature [23]. Thus, the above protocol is a direct application of the weak p-signature scheme proposed in [23], proven secure, and having properties: Signer Privacy, User privacy, Correctness, Unforgeability, and Zero-Knowledge.

Intuitively, security in the above protocol concerns two entities $\mathcal{KGC}$ and user. Following the same analysis in [10] the above protocol is secure if the underlying p-signature scheme is secure and the Signer Privacy, User privacy properties hold. In one hand, Signer Privacy ensures that a malicious user interacting with the $\mathcal{KGC}$ can't get any information on $\mathcal{KGC}$ master secret key other than user private key. On the other hand, the certificate presented by

the user to a malicious KGC reveals no information about the real identity of the user.

VI. Distributed anonymous SK-IBE

A. AKI for distributed SK-IBE

Smart el al. [11] presented a distributed version of Sakai-Kasahara scheme, in this section we give a modification of this scheme combined with the IND-ID-CCA scheme from [18], which we call DSK-IBE, assuring user anonymity when generating his private key by $\mathcal{KGC}$. As in Chow [10], master key generation is separated from the Setup stage, reducing further trust required in the $\mathcal{KGC}$.

DSK-IBE Setup : We first define explicitly the system's public parameters. Let $\mathbb{G}_1$, $\mathbb{G}_2$ and $\mathbb{G}_T$ denote groups of large prime order q, which are equipped with a bilinear pairing, $e : \mathbb{G}_1$ x $\mathbb{G}_2 \rightarrow \mathbb{G}_T$. We assume that $\mathbb{G}_1$, $\mathbb{G}_2$ are respectively generated by g and $\hat{g}$. We define four hash functions, $H_1 : \{0,1\}^* \rightarrow \mathbb{Z}_q^*$, $H_2 : \mathbb{G}_T \rightarrow \{0,1\}^n$, $H_3 : \{0,1\}^n$x$\{0,1\}^n \rightarrow \mathbb{Z}_q^*$ and $H_4 : \{0,1\}^n \rightarrow \{0,1\}^n$ for $n > 0$.

DSK-IBE KeyGen : distributed protocol runs between the n $\mathcal{KGC}$ and results in each $\mathcal{KGC}$ obtaining a share $s_i \in \mathbb{Z}_q$ of the master secret s. The tuple $\mathcal{C}_{\langle g \rangle}^{(s)} = [g^s, g^{s_1}, ..., g^{s_n}]$ is the system public-key. Note that a coalition of upto t entities should gain nothing about s, whereas $t+1$ entities could reconstruct the secret s.

ObtainKey$(\mathcal{U}(params, id, cert_U, open))$ $\leftrightarrow$ **IssueKey** $(\mathcal{KGC}_i(params, s_i, cert_U))$ **:** DAKI protocol run between m $\mathcal{KGC}$s ($t < m \leq n$) to produce m outputs $d_{id}^{(i)}$ which are shares of private key d_{id}. We modify the distributed SK-IBE Private key extraction [18] by adding steps 1 to 3 which made the protocol anonymous as follows:

1) User presents his certificate $cert_U$ to each one of the $t+1$ $\mathcal{KGC}$, the latter verifies certificate signature using the CA public key pk_{cert} if certificate verification fail it aborts the protocol.
2) The user chooses at random ρ in $\mathbb{Z}_q$;
3) The user and KGC_i (for $i = 1, ..., t$) engage in a secure two-party computational protocol [24], where the user's private input is $(\rho, H_1(ID), open)$, and the KGC_i's private input is s_i. As result, the KGC_i gets a private output which is either $S_i^{ID} = (s_i + H_1(ID))\rho$ if $com = commit(H_1(ID), open)$ or $\perp$ in this case the $\mathcal{KGC}_i$ aborts.
4) $\mathcal{KGC}_i$ runs $\left(\mathcal{C}_{\langle \hat{g}, \hat{h} \rangle}^{(z,z')}, z_i, z_i'\right) = Random_{Ped}(n, t, \hat{g}, \hat{h})$, where $\hat{h} \in \mathbb{G}_2$ is a generator for Pedersen commitments precomputed by $\mathcal{KGC}$s using $\left(\mathcal{C}_{\langle \hat{g} \rangle}^{(r)}\right) = Random_{DLog}(n, t, \hat{g})$, and set $\hat{h} = \left(\mathcal{C}_{\langle \hat{g} \rangle}^{(r)}\right)_0 = \hat{g}^r$. $\mathcal{KGC}_i$ also computes $\left(\mathcal{C}_{\langle g \rangle}^{(S_i^{ID})}\right)_j = g^{(s_j + H_1(ID))\rho}$ for $0 \leqslant j \leqslant n$.
5) $\mathcal{KGC}_i$ runs $\left(\mathcal{C}_{\langle \hat{g}, \hat{h} \rangle}^{(w,w')}, w_i, w_i'\right) = Mul_{Ped}(n, t, \hat{g}, \hat{h}, \left(\mathcal{C}_{\langle g \rangle}^{(S_i^{ID})}, S_i^{ID}\right), \left(\mathcal{C}_{\langle \hat{g}, \hat{h} \rangle}^{(z,z')}, z_i, z_i'\right))$, where $w = s^{ID}z = (s + H_{1(ID)})\rho z$, $w' = s^{ID}z' = (s + H_{1(ID)})\rho z'$ and sends $\left(\mathcal{C}_{\langle \hat{g}, \hat{h} \rangle}^{(w,w')}, w_i\right)$ along with $PK_1^{(i)} = NIZKPK_{\equiv} com(w_i, w_i', \left(\mathcal{C}_{\langle \hat{g}, \rangle}^{(w)}\right)_i, \left(\mathcal{C}_{\langle \hat{g}, \hat{h} \rangle}^{(w,w')}\right)_i)$ to the user.
6) $\mathcal{KGC}_i$ sends to the user $\left(\mathcal{C}_{\langle \hat{g} \rangle}^{(z)}\right)_i = \hat{g}^{z_i}$ and $\mathcal{C}_{\langle \hat{g}, \hat{h} \rangle}^{(z,z')}$ along with $PK_2^{(i)} = NIZKPK_{\equiv} com(z_i, z_i', \left(\mathcal{C}_{\langle \hat{g}, \rangle}^{(z)}\right)_i, \left(\mathcal{C}_{\langle \hat{g}, \hat{h} \rangle}^{(z,z')}\right)_i)$.

ReconstructKey$(w_i, \hat{g}^{z_i}, PK_1^{(i)}, PK_2^{(i)})$**:** Upon receiving $(w_i, \hat{g}^{z_i}, PK_1^{(i)}, PK_2^{(i)})$ for $(i = 1, ..., t+1)$ the user do the following computations:

1) verifies $\left(\mathcal{C}_{\langle \hat{g} \rangle}^{(z)}\right)_i$ using $PK_2^{(i)}$;
2) reconstructs (w, g^z) using Lagrange interpolation;
3) if $w = 0$ it aborts else it computes $w^{-1} = \frac{1}{(s+H_1(ID))\rho z}$;
4) computes his private key by: $d_{id} = (\hat{g}^z)^{w^{-1}\rho} = \left(\hat{g}^{\frac{1}{(x+H_1(ID))\rho}}\right)^{\rho} = \hat{g}^{\frac{1}{x+H_1(ID)}}$

B. Analysis

The above (n, t) SK-IBE scheme, without the anonymity propriety, was proven IND-ID-CCA secure in [18], assuming a standard t-limited Byzantine adversary in a system with n nodes, where any t nodes are compromised by the adversary. In contrast of this, and by adding steps (1) to (3), the obtained protocol is a distributed form of the p-signature scheme given in [23]. We argue that the new anonymous (n, t) SK-IBE scheme is IND-ID-CCA secure assuming the three following statements: (1) (n, t) SK-IBE scheme in [18] is IND-ID-CCA, (2) the underlying distributed p-signature scheme is unforgeable, satisfies issuer and user privacy, (3) the signature scheme used by $\mathcal{CA}$ is unforgeable.

Two primitives in the above construction are concerned by this analysis. Firstly, p-signature security follows the same rules as in Section (V.B) and results on user privacy, where a malicious $\mathcal{KGC}$ can't get any information about the real identity id of the user contained in the certificate $cert_U$ (this is ensured by the $\mathcal{CA}$ signature scheme proprieties that signs on a strongly computationally hiding commitment of id). For $\mathcal{KGC}$s privacy, due to the proprieties of the underlying secret sharing scheme, a malicious user can't get any information about $\mathcal{KGC}$ partial private key because shares he obtain $(w_i, \hat{g}^{z_i})$ reveal no information. Secondly, the signing algorithm of the certification authority is not specified, so the use of an unforgeable signature scheme, that signs on a perfectly binding and strongly computationally hiding commitment of the identity included in the certificate, should be fine for our purpose of security.

 ISBN:978-1-61208-113-7

VII. Conclusion

In this paper, we proposed an architecture for developing new class of distributed key issuing protocols that have the privacy-preserving propriety. Assuming this architecture, we proposed a new anonymous key issuing protocol for the distributed SK-IBE, which belongs to the exponent-inversion family, along with a informal security analysis.

Our construction is based on the recently proposed distributed private key generator [18], [11] combined with the anonymous key issuing protocol [10] thus coupling advantages of the two approaches. The proposed protocol aims to solve key escrow problem and single point of failure in IBE systems, which will reduce trust needed in $\mathcal{KGC}$.

Further work is required to extend the proposed protocol to other IBE frameworks (i.e. commutative-blinding IBEs and full-domain-hash IBEs), and to define a formal security model.

References

[1] A. Shamir, "Identity-based cryptosystems and signature schemes," in *CRYPTO 84*, 1985, pp. 47–53.

[2] C. Cocks, "An identity based encryption scheme based on quadratic residues," in *8th IMA International Conference*, ser. volume 2260 of LNCS. Springer Berlin, 2001, pp. 360–363.

[3] D. Boneh and M. Franklin, "Identity-based encryption from the weil pairing," in *Proceedings of the 21st Annual International Cryptology Conference on Advances in Cryptology*, ser. volume 2139 of LNCS. Springer Berlin, 2001, pp. 213–229.

[4] Y. Desmedt and M. Burmester, "Identity-based key infrastructures iki," in *SEC 2004*, 2004, pp. 167–176.

[5] L. Chen, K. Harrison, A. Moss, D. Soldera, and N. Smart, "Certification of public keys within an identity based system," in *ISC 2002*, ser. volume 2433 of LNCS, Springer-Verlag. Canterbury, UK, June 30 July 1, 2005, pp. 322–333.

[6] G. Price and C. J. Mitchell, "Interoperation between a conventional pki and an id-based infrastructure," in *EuroPKI 2005*, ser. volume 3545 of LNCS. Canterbury, UK, June 30 July 1, 2005, pp. 73–85.

[7] B. Lee, C. Boyd, E. Dawson, K. Kim, J. Yang, and S. Yoo, "Secure key issuing in id-based cryptography," in *proceedings of the Second Australian Information Security Workshop-AISW 2004, ACS Conferences in Research and Practice in Information Technology vol.32*, 2004, pp. 69–74.

[8] C. Gentry, "Certificate-based encryption and the certificate revocation problem," in *Advances in Cryptology - EUROCRYPT 2003*, ser. volume 2653 of LNCS. Springer Berlin, 2003, pp. 272–293.

[9] S. Al-Riyami and S. Paterson, "Certificateless public key cryptography," vol. 2894. Springer-Verlag, 2003, pp. 375–391.

[10] S. Chow, "Removing escrow from identity-based encryption," in *12th International Conference on Practice and Theory in Public Key Cryptography*, 2009, pp. 256–272.

[11] M. Geisler and N. P. Smart, "Distributing the key distribution centre in sakai-kasahara based systems," in *Conf. on Cryptography and Coding 09*, 2009, pp. 252–262.

[12] A. Kate and I. Goldberg, "A distributed private-key generator for identity-based cryptography, cryptology eprint archive, report 2009/355," Tech. Rep., 2009.

[13] L. Chen, K. Harrison, N. P. Smart, and D. Soldera, "Infrasec 2002," in *Applications of multiple trust authorities in pairing based cryptosystems*, ser. volume 2437 of LNCS, Springer-Verlag, 2002, pp. 260–275.

[14] K. Paterson, "Cryptography from pairings: a snap shot of current research," Information Security Technical Report 7 (3), Tech. Rep., 2002.

[15] F. Hess, "Efficient identity based signature schemes based on pairings," in *Selected Areas in CryptographySAC 02*, ser. volume 2595 of LNCS, Springer-Verlag, 2002, pp. 310–324.

[16] S. Kwon, "Cryptanalysis for secure key issuing in id-based cryptography and improvement, manuscript," Tech. Rep., 2004.

[17] B. Lee, E. Dawson, and S. Moon, "Efficient and robust secure key issuing in id-based cryptography," in *proceedings of the 6-th International Workshop on Information Security Applications (WISA 2005)*, 2005, pp. 267–280,.

[18] A. Kate and I. Goldberg, "Distributed private-key generators for identity based cryptography," in *To appear at SCN 10*, 2010.

[19] A. Sui, S. S. M. Chow, L. C. K. Hui, S. M. Yiu, K. P. Chow, W. W. Tsang, C. F. Chong, K. H. Pun, and H. W. Chan, "Seperable and anonymous identity-based key issuing without secure channel," in *11th International Comference on Parallel and Distributed Systems ICPADS*, ser. volume 2, 2005, pp. 275–279.

[20] T. P. Pedersen, "Non-interactive and information-theoretic secure verifiable secret sharing," in *Cryptology CRYPTO 91*, 1991, pp. 129–140.

[21] M. Abdalla, M. Bellare, D. Catalano, E. Kiltz, T. Kohno, T. Lange, J. Malone-lee, G. Neven, P. Paillier, and H. Shi, "Searchable encryption revisited: Consistency properties, relation to anonymous ibe, and extensions," in *In CRYPTO*. Springer-Verlag, 2005, pp. 205–222.

[22] M. Izabachene and D. Pointcheval, "New anonymity notions for identity-based encryption," in *the 6th international conference on Security and Cryptography for Networks*, 2008, pp. 375–391.

[23] M. Belenkiy, M. Chase, M. Kohlweiss, and A. Lysyanskaya, "P-signatures and noninteractive anonymous credentials," in *In Theory of Cryptography Conference*, 2008, pp. 356–374.

[24] S. Jarecki and V. Shmatikov, "Efficient two-party secure computation on committed inputs," in *EUROCRYPT 07*, 2007, pp. 97–114.

 ISBN:978-1-61208-113-7

Efficiency Optimisation Of Tor Using Diffie-Hellman Chain

Kun Peng
Institute for Infocomm Research, Singapore
dr.kun.peng@gmail.com

***Abstract*—Onion routing is the most common anonymous communication channel. Usually onion routing is specified through asymmetric cipher and thus is inefficient. In Tor (the second generation onion router), it is suggested to employ symmetric cipher to encrypt the packets in onion routing. Obviously, symmetric cipher is much more efficient than the asymmetric cipher employed in the original onion routing. However, whether this idea can really work depends on whether an efficient (both in computation and communication) key generation and exchange mechanism can be designed for the symmetric cipher to employ. The suggestion in Tor is simple and it is a direct employment of Diffie-Hellman handshake to generate the secret keys for the routers' symmetric cipher. In this paper we show that direct application of Diffie-Hellman handshake to implement key generation and exchange in onion routing is not efficient in communication as multiple instances of Diffie-Hellman handshake needs a lot of additional communication. Moreover, its efficiency improvement for the sender is not satisfactory. So we design a more advanced application of Diffie-Hellman key exchange technique, Diffie-Hellman chain. This new technique greatly saves a sender's cost and needs very few communication for Diffie-Hellman key exchange. With the efficiency improvement in this paper, Tor can be applied to communication networks with weaker computational capability and smaller communicational bandwidth.**

Index Terms*—TOR; efficient key exchange; Diffie-Hellman chain

I. INTRODUCTION

Anonymous communication channel is a very useful tool in e-commerce, e-government and other cryptographic applications, which often require anonymity and privacy. In an anonymous communication channel, the messages are untraceable, so can be transmitted anonymously. A common method to implement anonymous channels is onion routing [1], [3], [4], which employs multiple nodes to route a message. A node in an onion routing communication network can send a message to any node in the network. The sender can flexibly choose any route from all the connection paths between him and the receiver. Each message is contained in a packet called an onion. In the packet, a message is encrypted layer by layer using the encryption keys of all the routers on its route and the receiver. Each layer of encryption is just like a layer of onion bulb. In onion routing, given a message packet, each router unwraps a layer of encryption by decrypting the message packet using its decryption key, finds out the identity of the next router and forwards the unwrapped message packet to the next router. Unless gaining collusion of all the routers on the routing path of his received message, the receiver cannot trace the message back to the sender, who then obtains anonymity. When a packet is routed together with a large number of other packets, onion routing prevents it from being traced, even if the whole onion network is monitored.

An obvious advantage of onion routing over other specifications of anonymous communication channel (e.g. mix network [5], [6], which sends multiple messages from a unique sender to a unique receiver through a unique path) is that each of multiple senders can send his message to any of multiple receivers and freely choose a dynamic routing path and so higher flexibility and applicability are achieved. Another advantage of onion routing will be illustrated in this paper in our new routing protocols: feasibility to get rid of costly asymmetric encryption and decryption, which are inevitable in mix networks.

The key technique in onion routing is encryption chain, in which a message is successively encrypted with multiple keys. More precisely, multiple keys form a chain and are employed one by one to encrypt a message. Not only the message, the identity of each router on its routing path is encrypted in an encryption chain using the encryption keys of all the routers before it. When an onion packet is routed, each router unwraps it by removing one layer of encryption from each encryption chain. So each router can recover the identity of the next router and forward the partially decrypted packet. In onion routing, the encryption chains are usually implemented through asymmetric cipher. Namely, the message and identities of the routers are encrypted using the routers' public keys and the routers unwrap the onion packet using their private keys. An advantage of using asymmetric cipher is that with the support of PKI or ID-based public key system no special key exchange operation is needed. As there are multiple encryption chains (one for the message and one for each router) and the there are $O(n^2)$ encryption and decryption operations (where n is the number of routers), such an implementation through asymmetric cipher is inefficient.

Tor [2] is the second generation of onion routing. It proposes a few optimisations for onion routing. A suggested optimisation in Tor is to replace asymmetric cipher with much more efficient symmetric cipher to improve efficiency of onion routing. It is a commom sense that symmetric cipher is much more efficient than asymmetric cipher. The key point in using symmetric cipher is how to distribute the session keys using public key operations, while a simple solution to the key-exchange problem in application of symmetric cipher is the

 ISBN:978-1-61208-113-7

Diffie-Hellman key exchange protocol recalled in Section II-B. So it is suggested in Tor [2] to employ "Diffie-Hellman handshake" to implement key changes and generate session keys for the roueters. As the idea is only simply mentioned and not specified in details in [2], it is specified in Section IV in this paper to assess its effect. Our assessment illustrates that although improving computational efficiency of the routers the suggested efficiency improvement in Tor [2] is not very satisfactory. Firstly, it greatly increases communicational cost. Secondly, even using it the sender's computational cost is still high.

The symmetric-cipher-based key chain in Tor [2] is optimised in this paper. Firstly, we optimise the "Diffie-Hellman handshake" and reduce the number of communication rounds in Tor and obtain a simple optimisation. As it is still a direct application of Diffie-Hellman key exchange, its efficiency improvement is still not satisfactory. So Diffie-Hellman key exchange is then extended and adapted for onion routing in a more advanced way such that a sender can efficiently distribute the sysmmetric sessions keys to the routers through the onion packet. The new key exchange technique is called Diffie-Hellman chain, which chain up the Diffie-Hellman handshakes for the routers and receiver such that they are much more efficient then separate Diffie-Hellman handshakes. An efficient onion routing protocol is designed in Section V using Diffie-Hellman chain. It employs Diffie-Hellman chain and block cipher encryption chain to improve computational and communicational efficiency of Tor. The new onion routing protocol is more appliable than most onion routing implementations including Tor. Network with smaller bandwidth and lower-power routers can employ them to achieve anonymity.

II. Preliminaries

Symbol denotions and background knowledge to be used in this paper are introduced and recalled in this section.

A. Parameter Setting and Symbols

The following symbols are used in this paper.

- p and q are large primes and q is a factor of $p-1$. G is the cyclic subgroup with order q in Z_p^*. g is a generator of G.
- Encryption of m using key k is denoted as $E_k(m)$ where block cipher (e.g. AES) is employed.
- Encryption chain of m using block cipher and key $k_1, k_2, \ldots, k_i$ is denoted as $E_{k_1,k_2,\ldots,k_i}(m)$. The encryptions are performed layer by layer. k_1 is the the key used in the most outer layer; k_2 is the the key used in the second most outer layer; $\ldots$; k_i is the the key used in the most inner layer.
- In onion routing, the routers are $P_1, P_2, \ldots, P_n$ and the receiver is denoted as the last router P_{n+1}.
- The private key of P_i is x_i, which is randomly chosen from Z_q. The corresponding public keys are $y_1, y_2, \ldots, y_n$ where $y_i = g^{x_i} \bmod p$ for $i = 1, 2, \ldots, n$.

B. Diffie-Hellman Key Exchange

Symmetric ciphers like block cipher are very efficient. However, unlike asymmetric cipher they depend on key exchange protocols to distribute keys. The most common key exchange protocol is Diffie-Hellman key exchange protocol. Two parties A and B can cooperate to generate a session key as follows.

1) A randomly chooses α from Z_q and sends his key base $\mu = g^\alpha \bmod p$ to B.
2) B randomly chooses β from Z_q and sends his key base $\nu = g^\beta \bmod p$ to A.
3) A can calculate the key $k = \nu^\alpha \bmod p$, while B can calculate the key $k = \mu^\beta \bmod p$.

The famous Diffie-Hellman problem is recalled as follows.

Definition 1: (Diffie-Hellman problem) Given μ and ν, it is difficult to calculate k if the discrete logarithm problem is hard.

III. Specifying and Assessing the Suggested Efficiency Improvement in Tor

The suggestion to employ symmetric cipher in Tor [2] is quite simple. To precisely assessing its cost and comparing it with our new design of key exchange, we need to specify it in details. For simplicity of description, our specification focuses on efficiency improvement through symmetric cipher as it is the focus of this paper, while the other optimisations of onion routing in Tor are ignored. The suggested efficiency improvement in Tor is specified in details as follows where a message m is sent by a sender through n routers $P_1, P_2, \ldots, P_n$ to a receiver P_{n+1}.

1) For the receiver and each router P_i where $1 \le i \le n+1$, the sender randomly chooses an integer s_i from Z_q and calculates $\hat{k}_i = g^{s_i} \bmod p$.
2) The sender sends $\hat{k}_1$ to P_1, which returns $\hat{k}'_1 = g^{s'_1} \bmod p$ where s'_1 is randomly chosen from Z_q. Both the sender and P_1 obtains their session key $k_1 = g^{s_1 s'_1} \bmod p$.
3) The sender sends $E_{k_1}(P_2)$ and $E_{k_1}(\hat{k}_2)$ to P_1, who decrypts the two ciphertexts using his session key and then sends $\hat{k}_2$ and $\hat{k}'_1$ to P_2.
4) P_2 randomly chooses s'_2 from Z_q and obtains his session key with the sender $k_2 = \hat{k}_2^{s'_2} = g^{s_2 s'_2}$ and his session key with P_1, $K_{1,2} =$. He sends E
5) The sender encrypts the message m, the key base list $g^{s_1}, g^{s_2}, \ldots, g^{s_{n+1}}$ and the route list $p_1, p_2, \ldots, p_{n+1}$ as follows.
   a) He calculates $e = E_{k_1,k_2,\ldots,k_{n+1}}(m)$.
   b) He calculates $K_i = E_{k_1,k_2,\ldots,k_{i-1}}(g^{s_i})$ for $i = 1, 2, \ldots, n+1$.
   c) He calculates $p_i = E_{k_1,k_2,\ldots,k_i}(P_{i+1})$ for $i = 1, 2, \ldots, n+1$ where $P_{n+2} = P_{n+1}$.
   d) He sends out the initial onion

$$O_1 = (a_1, b_{1,1}, b_{1,2}, \ldots, b_{1,n+1}, c_{1,1}, c_{1,2}, \ldots, c_{1,n+1}) = (e, K_1, K_2, \ldots, K_{n+1}, p_1, p_2, \ldots, p_{n+1})$$

to P_1.

6) Each router P_i routes the onion as follows where the onion is in the form $O_i = (a_i, b_{i,1}, b_{i,2}, \ldots, b_{i,n+1}, c_{i,1}, c_{i,2}, \ldots, c_{i,n+1})$ when it is sent to P_i.
   a) P_i generates his session key $k_i = b_{i,1}^{x_i} \bmod p$.
   b) P_i uses k_i to decrypt $c_{i,j}$ for $j = 1, 2, \ldots, n+1$ and obtains $P_{i+1} = D_{k_i}(c_{i,1})$.
   c) P_i uses k_i to decrypt a_i and obtains $a_{i+1} = D_{k_i}(a_i)$.
   d) Finally, P_i sends
   $$O_{i+1} = (a_{i+1}, b_{i+1,1}, b_{i+1,2}, \ldots, b_{i+1,n+1}, c_{i+1,1}, c_{i+1,2}, \ldots, c_{i+1,n+1})$$
   to P_{i+1} where $b_{i+1,j} = D_{k_i}(b_{i,j+1})$ and $c_{i+1,j} = D_{k_i}(c_{i,j+1})$ for $j = 1, 2, \ldots, n$ and $b_{i+1,n+1}$ and $c_{i+1,n+1}$ are two random integers in the ciphertext space of the employed symmetric encryption algorithm.
7) At last, P_{n+1} receives
   $$O_{n+1} = (a_{n+1}, b_{n+1,1}, b_{n+1,2}, \ldots, b_{n+1,n+1}, c_{n+1,1}, c_{n+1,2}, \ldots, c_{n+1,n+1})$$
   and operates as follows.
   a) P_{n+1} generates his session key $k_{n+1} = b_{n+1,1}^{x_{n+1}} \bmod p$.
   b) P_{n+1} uses k_{n+1} to decrypt $c_{n+1,j}$ and obtains $P_{n+1} = D_{k_{n+1}}(c_{n+1,1})$.
   c) P_{n+1} knows that itself is the receiver as P_{n+1} is its own identity.
   d) P_{n+1} uses k_{n+1} to decrypt a_{n+1} and obtains $m = D_{k_{n+1}}(a_{n+1})$.

IV. A Simple Optimisation of Tor and its Drawback: Simpler but still Direct Application of Diffie-Hellman Key Exchange

A simple optimisation of Tor is proposed in this section. Like in the original onion routing (and many other cryptographia protocols), it assumes that every router and the receiver have discrete-logarithm-based public key encryption algorithms (e.g. ElGamal encryption) and already set up their public keys so that half of the preparation work in Diffie-Hellman key exchange can be saved. Moreover, multiple rounds of communication between each pair of participants are combined to improve communication efficiency. It still employ Diffie-Hellman handshakes in the staightforward way and is described as follows.

1) For the receiver and each router P_i where $1 \leq i \leq n+1$, the sender randomly chooses an integer s_i from Z_q and generates a session key $k_i = y_i^{s_i}$.
2) The sender encrypts the message m, the key base list $g^{s_1}, g^{s_2}, \ldots, g^{s_{n+1}}$ and the route list $p_1, p_2, \ldots, p_{n+1}$ as follows.
   a) He calculates $e = E_{k_1,k_2,\ldots,k_{n+1}}(m)$.
   b) He calculates $K_i = E_{k_1,k_2,\ldots,k_{i-1}}(g^{s_i})$ for $i = 1, 2, \ldots, n+1$.
   c) He calculates $p_i = E_{k_1,k_2,\ldots,k_i}(P_{i+1})$ for $i = 1, 2, \ldots, n+1$ where $P_{n+2} = P_{n+1}$.
   d) He sends out the initial onion
   $$O_1 = (a_1, b_{1,1}, b_{1,2}, \ldots, b_{1,n+1}, c_{1,1}, c_{1,2}, \ldots, c_{1,n+1}) = (e, K_1, K_2, \ldots, K_{n+1}, p_1, p_2, \ldots, p_{n+1})$$
   to P_1.
3) Each router P_i routes the onion as follows where the onion is in the form $O_i = (a_i, b_{i,1}, b_{i,2}, \ldots, b_{i,n+1}, c_{i,1}, c_{i,2}, \ldots, c_{i,n+1})$ when it is sent to P_i.
   a) P_i generates his session key $k_i = b_{i,1}^{x_i} \bmod p$.
   b) P_i uses k_i to decrypt $c_{i,j}$ for $j = 1, 2, \ldots, n+1$ and obtains $P_{i+1} = D_{k_i}(c_{i,1})$.
   c) P_i uses k_i to decrypt a_i and obtains $a_{i+1} = D_{k_i}(a_i)$.
   d) Finally, P_i sends
   $$O_{i+1} = (a_{i+1}, b_{i+1,1}, b_{i+1,2}, \ldots, b_{i+1,n+1}, c_{i+1,1}, c_{i+1,2}, \ldots, c_{i+1,n+1})$$
   to P_{i+1} where $b_{i+1,j} = D_{k_i}(b_{i,j+1})$ and $c_{i+1,j} = D_{k_i}(c_{i,j+1})$ for $j = 1, 2, \ldots, n$ and $b_{i+1,n+1}$ and $c_{i+1,n+1}$ are two random integers in the ciphertext space of the employed symmetric encryption algorithm.
4) At last, P_{n+1} receives
   $$O_{n+1} = (a_{n+1}, b_{n+1,1}, b_{n+1,2}, \ldots, b_{n+1,n+1}, c_{n+1,1}, c_{n+1,2}, \ldots, c_{n+1,n+1})$$
   and operates as follows.
   a) P_{n+1} generates his session key $k_{n+1} = b_{n+1,1}^{x_{n+1}} \bmod p$.
   b) P_{n+1} uses k_{n+1} to decrypt $c_{n+1,j}$ and obtains $P_{n+1} = D_{k_{n+1}}(c_{n+1,1})$.
   c) P_{n+1} knows that itself is the receiver as P_{n+1} is its own identity.
   d) P_{n+1} uses k_{n+1} to decrypt a_{n+1} and obtains $m = D_{k_{n+1}}(a_{n+1})$.

This modified Tor protocol only employs symmetric cipher in encryption and decryption operations. The only public key operations in it are $n+1$ instances of Diffie-Hellman key exchange. So although more encryption and decryption operations are needed than in traditional onion routing, it is still more efficient in computation. However, it is less efficient in communication than traditional onion routing as its onion packet contains additional encrypted key bases $b_{i,1}, b_{i,2}, \ldots, b_{i,n+1}$. So its advantage in efficiency is not obvious. Therefore, it is only a prototype, while our final proposal is based on it but has higher requirements on efficiency: only using symmetric cipher in encryption and decryption while in comparison with traditional onion routing

 ISBN:978-1-61208-113-7

- very little additional communication (e.g. one more integer) is needed;
- no more additional encryption or decryption operation is needed.

V. A New and More Advanced Technique: Diffie-Hellman Chain

The simple optimisation protocol in Section IV has demonstrated that direct application of Diffie-Hellman key exchange to onion routing (including original onion routing and Tor) cannot achieve satisfiactory advantage in efficiency. To reduce additional communication and encryption and decryption operations, a novel technique, Diffie-Hellman chain, is designed. The Diffie-Hellman key bases for all the routers and the receiver are sealed in the Diffie-Hellman chain, which appears in each onion packet in the form of a single integer. For each router, to generate his session key, he needs his private key and a key base initially sealed in the Diffie-Hellman chain by the sender and then recovered by cooperation of all the previous routers in the course of routing. As only one single integer is needed in each onion packet to represent the Diffie-Hellman chain and commit to all the Diffie-Hellman key bases, a very small amount of additional communication is employed and no more encryption (decryption) operation is needed in comparison with traditional onion routing.

A new onion routing protocol, called compressed onion routing, is proposed. In compressed onion routing, a packet (onion) consists of three parts: message, route list and key base. Route list contains the identities of all the nodes on the route. Key base is the base to generate the session keys (symmetric keys) distributed to the nodes. The message part in compressed onion routing is similar to that in most onion routing schemes. The message is encrypted in a encryption chain using the sessions keys of all the nodes. The readers only need to note that efficient block cipher is employed in the encryption chain. In compressed onion routing, the route list is the same as in other onion routing schemes. It consists of all the routers' identities. One encryption chain is used to seal each router's identity using the session keys of the all the routers before it. The readers only need to note that efficient block cipher is employed in the encryption chains for the route list.

The most important novel technique is generation and update of the key base, which enables key exchange. Each router builds his session key on the base of the key base using his private key and update the key base for the next router. The key generation function is similar to Diffie-Hellman key generation, but we do not employ separate Diffie-Hellman key exchange protocols to distribute the session keys to the routers. Instead the key base updating mechanism actually generates a key base chain and so all the session keys and their generation functions are linked in a chain structure. So the key exchange technique is called Diffie-Hellman chain. After obtaining his session key, each router can extracts the identity of the next router from the route list using his session key, removes one layer of encryption from the message using his session key and then forwards the onion to the next router. The Diffie-Hellman chain only needs the bandwidth of one integer, and thus is much more efficient than separate key distribution in communication. Novelty of the new compressed onion routing protocol is that distribution of the sessions keys and encryption of the routers' identities are compressed such that fewer computationally-costly public key operations and communicationally-costly encryption chains are needed.

Suppose a message m is sent by a sender through n routers $P_1, P_2, \ldots, P_n$ to the receiver P_{n+1}. Firstly, the sender generates the session keys $k_1, k_2, \ldots, k_{n+1}$ respectively for $P_1, P_2, \ldots, P_{n+1}$ as follows.

1) The sender randomly chooses an integer s_1 from Z_q.
2) The sender calculates P_1's session key $k_1 = y_1^{s_1} \bmod p$.
3) The sender calculates $s_2 = s_1 + k_1 \bmod q$.
4) The sender calculates P_2's session key $k_2 = y_2^{s_2} \bmod p$.
5) ……
6) ……
7) The sender calculates $s_{n+1} = s_n + k_n \bmod q$.
8) The sender calculates P_{n+1}'s session key $k_{n+1} = y_{n+1}^{s_{n+1}} \bmod p$.

Generally speaking, for $i = 1, 2, \ldots, n+1$, the sender

1) if $i > 1$ then calculates $s_i = s_{i-1} + k_{i-1} \bmod q$ as his secret seed in the Diffie-Hellman chain for generation of k_i
2) calculates $k_i = y_i^{s_i}$

where s_1 is randomly chosen from Z_q. In summary, the sender uses the sum of the previous node's session key and his secret seed in the Diffie-Hellman generation of the previous node's session key as his secret seed to generate a node's Diffie-Hellman session key. The other secret seed to generate the node's session key is the node's private key.

The route list consists of $p_1, p_2, \ldots, p_{n+1}$ where $p_i = E_{k_1,k_2,\ldots,k_i}(P_{i+1})$ and $P_{n+2} = P_{n+1}$. The message is encrypted into $e = E_{k_1,k_2,\ldots,k_{n+1}}(m)$. The onion is in the form of $O_i = (a_i, b_i, c_{i,1}, c_{i,2}, \ldots, c_{i,n+1})$ when it reaches P_i where a_i is the encrypted message, b_i is the key base and $c_{i,1}, c_{i,2}, \ldots, c_{i,n+1}$ is the encrypted route list. Note that although the encryption chain for the next router's identity is completely decrypted and discarded by each router, the length of the encrypted route list is kept unchanged for the sake of untraceability. If an onion packet becomes shorter after each router's routing, its change in length can be observed and exploited to trace it. So we keep the length of each encrypted route list constant to maintain the size of onion packets. This can be implemented by inserting a random tag into the onion packets after they discard an encryption chain. The initial onion $O_1 = (a_1, b_1, c_{1,1}, c_{1,2}, \ldots, c_{1,n+1}) = (e, g^{s_1}, p_1, p_2, \ldots, p_{n+1})$. Note that e may actually contain multiple symmetric ciphertext blocks as the message may be long and is divided into multiple blocks when being encrypted. For convenience of description encryption of the message is still denoted as a single variable and the readers should be aware that it is the encryption of the whole message and may contain multiple blocks.

 ISBN:978-1-61208-113-7

P_1 receives $O_1 = (a_1, b_1, c_{1,1}, c_{1,2}, \ldots, c_{1,n+1})$ from the sender and then operates as follows.

1) P_1 generates his session key $k_1 = b_1^{x_1} \bmod p$.
2) P_1 uses k_1 to decrypt $c_{1,j}$ for $j = 1, 2, \ldots, n+1$ and obtains $P_2 = D_{k_1}(c_{1,1})$.
3) P_1 uses k_1 to decrypt a_1 and obtains $a_2 = D_{k_1}(a_1)$.
4) P_1 calculates the new key base $b_2 = b_1 g^{k_1} \bmod p$.

Finally, P_1 sends $O_2 = (a_2, b_2, c_{2,1}, c_{2,2}, \ldots, c_{2,n+1})$ to P_2 where $c_{2,i} = D_{k_1}(c_{1,i+1})$ for $i = 1, 2, \ldots, n$ and $c_{2,n+1}$ is a random integer in the ciphertext space of the employed block encryption algorithm.

More generally, for $i = 1, 2, \ldots, n$ each P_i receives $O_i = (a_i, b_i, c_{i,1}, c_{i,2}, \ldots, c_{i,n+1})$ and operates as follows.

1) P_i generates his session key $k_i = b_i^{x_i} \bmod p$.
2) P_i uses k_i to decrypt $c_{i,j}$ for $j = 1, 2, \ldots, n+1$ and obtains $P_{i+1} = D_{k_i}(c_{i,1})$.
3) P_i uses k_i to decrypt a_i and obtains $a_{i+1} = D_{k_i}(a_i)$.
4) P_i calculates the new key base $b_{i+1} = b_i g^{k_i} \bmod p$.

Finally, P_i sends $O_{i+1} = (a_{i+1}, b_{i+1}, c_{i+1,1}, c_{i+1,2}, \ldots, c_{i+1,n+1})$ to P_{i+1} where $c_{i+1,j} = D_{k_i}(c_{i,j+1})$ for $j = 1, 2, \ldots, n$ and $c_{i+1,n+1}$ is a random integer in the ciphertext space of the employed symmetric encryption algorithm.

At last, P_{n+1} receives $O_{n+1} = (a_{n+1}, b_{n+1}, c_{n+1,1}, c_{n+1,2}, \ldots, c_{n+1,n+1})$ and operates as follows.

1) P_{n+1} generates his session key $k_{n+1} = b_{n+1}^{x_{n+1}} \bmod p$.
2) P_{n+1} uses k_{n+1} to decrypt $c_{n+1,j}$ and obtains $P_{n+1} = D_{k_{n+1}}(c_{n+1,1})$.
3) P_{n+1} knows that itself is the receiver as P_{n+1} is its own identity.
4) P_{n+1} uses k_{n+1} to decrypt a_{n+1} and obtains $m = D_{k_{n+1}}(a_{n+1})$.

VI. Analysis and Comparison

Security of the compressed onion routing scheme depends on hardness of Diffie-Hellman problem as its key exchange mechanism is an extension of Diffie-Hellman key exchange. Its main trick is combining key exchange with encryption chain such that every router can obtain his session key with the help the previous router. As security of Diffie-Hellman key exchange has been formally proved and hardness of Diffie-Hellman problem is widely accepted, no further proof of security is needed except for Theorem 1, which shows that the session keys can be correctly exchanged.

Theorem 1: For $j = 1, 2, \ldots, n+1$, the same session key k_i is generated, respectively by the sender as $k_i = y_i^{s_i} \bmod p$ and by P_i as $k_i = b_i^{x_i} \bmod p$.

To prove Theorem 1, a lemma has to be proved first.

Lemma 1: For $j = 1, 2, \ldots, n+1$, $b_i = g^{s_i} \bmod p$.

Proof: Mathematical induction is used.

1) When $i = 1$, $b_1 = g^{s_1} \bmod p$
2) When $i = j$, suppose $b_j = g^{s_j} \bmod p$. Then a deduction can be made in next step.

TABLE I
Comparison of the anonymous communication channels

Scheme	public key exponentiation	flexibility and applicability
Mix network	$\geq 6n + 4$	No
AOS	$2(n+1)(n+4)$	Yes
Tor	$2(2n-1)$	Yes
COR	$3(n+1)$	Yes

TABLE II
Computational Efficiency Comparison for the Sender

Scheme	public key exponentiation	block cipher encryption
AOS	$(n+1)(n+4)$	0
Tor	$n+1$	$(n+1)(1+(3n+2)/2)$
COR	$n+1$	$(n+1)(1+(n+2)/2)$

3) When $i = j + 1$, $b_{j+1} = b_j g^{k_j} = g^{s_j} g^{k_j} \bmod p$ as it is supposed in last step that $b_i = g^{s_i}$ when $i = j$. So

$$b_{j+1} = g^{s_j} g^{k_j} = g^{s_j + k_j} = g^{s_{j+1}} \bmod p$$

Therefore, $b_i = g^{s_i} \bmod p$ for $j = 1, 2, \ldots, n+1$ as a result of mathematical induction. □

Proof of Theorem 1:
According to Lemma 1,

$$y_i^{s_i} = g^{x_i s_i} = b_i^{x_i} \bmod p$$

for $j = 1, 2, \ldots, n+1$. □

Efficiency comparison between our new onion routing protocol and the existing anonymous communication channels is given in Table I, Table II, Table III and Table IV where AOR stands for asymmetric cipher based onion routing and COR stands for compressed onion routing. The first table shows the advantage of our new technique over the existing anonymous communication channels including onion routing and mix network. The last three tables show our optimisation of onion routing. It is assumed that the employed block cipher is 256-bit AES. For simplicity, it is assumed that the message is one block long, while the size of one block of the employed block cipher should be large enough for a router's identity. So all the ciphertexts are one block long in our analysis, which does not lose generality and can be extended to long message cases in a straightforward way. As for asymmetric cipher in AOR, it is supposed that ElGamal encryption, which is the most popular with onion routing, is employed. More precisely, it is assumed that the ElGamal encryption algorithm uses 1024-bit integers. Comparison in the four tables illustrates that great efficiency improvement is achieved in the two compressed onion routing protocols.

VII. Conclusion

The new onion routing scheme proposed in this paper greatly improves efficiency of onion routing by using symmetric cipher and Diffie-Hellman chain. It needs smaller packet

TABLE III
COMPUTATIONAL EFFICIENCY COMPARISON FOR A ROUTER (RECEIVER)

Scheme	average public key exponentiation	average block cipher decryption
AOS	$n+4$	0
Tor	3	$2(n+1)$
COR	2	$(n+4)/2$

TABLE IV
COMMUNICATIONAL EFFICIENCY COMPARISON

Scheme	number of bits in an onion packet	rounds
AOS	$2048(n+2)$	$n+1$
Tor	$256(n+2)$	$(n+1)(n+3)$
COR	$256(n+2)+1024$	$n+1$

size and less computation than the existing onion routing schemes including TOR.

An open question in the future work is how to further compress the size of onion packets. The route list chains occupy most room in an onion packet. Can they be compressed to further improve communication efficiency?

REFERENCES

[1] J. Camenisch and A. Mityagin. A formal treatment of onion routing. In *CRYPTO '05*, volume 3089 of *Lecture Notes in Computer Science*, pages 169–187, Berlin, 2005. Springer-Verlag.

[2] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *USENIX Security Symposium*, pages 303–320, 2004.

[3] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game or a completeness theorem for protocols with honest majority. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing, STOC 1987*, pages 218–229, 1987.

[4] D. Goldschlag, M. Reed, and P. Syverson. Onion routing for anonymous and private internet connections. *Comm. of the ACM, 42(2)*, page 84–88, 1999.

[5] K. Peng, C. Boyd, and E. Dawson. Simple and efficient shuffling with provable correctness and ZK privacy. In *PKC '04*, volume 2947 of *Lecture Notes in Computer Science*, pages 439–454, Berlin, 2004. Springer-Verlag.

[6] K. Peng, C. Boyd, and E. Dawson. Simple and efficient shuffling with provable correctness and ZK privacy. In *CRYPTO '05*, volume 3089 of *Lecture Notes in Computer Science*, pages 188–204, Berlin, 2005. Springer-Verlag.

 ISBN:978-1-61208-113-7

Interaction between an Online Charging System and a Policy Server

Marc Cheboldaeff
Carrier Applications EMEA
Alcatel-Lucent
Ratingen, Germany
Marc.Cheboldaeff@alcatel-lucent.com

Abstract— **According to the 3GPP standard architecture up to Release 9, a Charging System is not supposed to interact directly with a Policy Server. The Charging System is responsible for rating and charging, while the Policy Server is responsible for determining the right policy depending on the kind of traffic. In reality, it appears that the decision about the right policy might be influenced by some real-time subscriber information, which might also be relevant for charging, and therefore stored in the Charging System. In this context, a direct interface between the Charging System and the Policy Server might be required. The goal of this paper is to study what such an interface would look like, based on an actual implementation. The main achievement is to validate a scenario where the policy should change in real-time during a data session because a volume threshold has been crossed.**

Keywords- Rating; Policy; IMS; OCS; PCRF; PCC; QoS

I. INTRODUCTION

According to the 3GPP IP Multimedia Subsystem (IMS) standard architecture, rating and charging takes place for online charging in a so-called Online Charging System (OCS) [1]. The latter contains rating and charging rules depending on all traffic typology criteria. The decision regarding policy falls to the Policy and Control Resource Function (PCRF) [2]. The latter contains policy rules depending on all traffic typology criteria.

In the standard Policy and Charging Control (PCC) architecture [3], the core network, which knows the actual traffic properties, asks the PCRF which policy it should apply, and the OCS (in case of online charging) which charging scheme it should apply. The Policy and Charging Enforcement Function (PCEF) at core network level, included in a Packet Data Network (PDN) Gateway, which can be for example the Gateway GPRS Support Node (GGSN), is then responsible to apply the proper policy and the correct pricing structure to the actual traffic, according to the input from the PCRF and the OCS.

However, there are some scenarios where the decision on policy might be influenced by the OCS. In Section II, we shall describe such scenarios, and check whether some of them have been studied already in the literature. In Section III, we will describe an approach in which the OCS interacts with the PCRF. Finally, in Section IV, we will present an actual implementation.

Please note that we focus on online charging in this paper, not on offline charging, because we are interested in rating respectively policy decisions / changes in real-time while a data session is running.

II. PURPOSE OF THE INTERACTION BETWEEN THE PCRF AND THE OCS

Going into more details in the IMS standard architecture for online charging [1], the OCS relies on two databases:

- The database in the Rating Function (RF), which contains generic tariff information at service level;
- The database in the Account Balance Management Function (ABMF), which contains subscriber-specific information relevant for the rating.

Actually, searching the literature, an interface between the policy decision function and external databases is mentioned in [4], but it does not relate specifically to an OCS database. And the dynamic mid-session interaction is not studied in detail either. A direct interaction between the PCRF and the OCS has already been studied in [5], but it restricts to an interaction of the PCRF with the Rating or Tariff Function of the OCS. It means that the policy decision might indeed depend on tariff rules, but it still does not depend on subscriber-specific information such as his/her current consumption or life cycle state.

Moreover, reducing the subscriber tariff information to a single tariff class ID might be restrictive given newer tariff schemes, where multiple charging options might be applied individually on top of a default tariff. Such charging options are for example usage-based discounts, subscriber bonus, or individual buckets e.g., free minutes, that the subscriber can book in addition to his/her default tariff, or that he/she gets as a reward for high consumption or recharge.

Basically, one of the functions of the OCS is to perform account balance management towards external systems through the ABMF. For this purpose, the OCS might store subscriber's pieces of information applicable for rating like usage counters. Furthermore, it might store additional information like his/her life-cycle state e.g., validity dates, or the status of his/her valid tariff options.

According to [1], in order to support the online rating process, the Rating Function necessitates counters. The counters are maintained by the Rating Function through the Account Balance Management Function. Assuming that these counters are maintained at subscriber level, storing them together with other real-time subscriber information in the ABMF makes sense.

 ISBN:978-1-61208-113-7

According to [3], in order to support the policy decision process, the PCRF may receive information about total allowed usage per user from the Subscription Profile Repository (SPR). Going further in this direction, some additional subscriber information might be relevant to the PCRF in order to determine the right policy: not only static data like an allowed usage threshold specific to a subscriber, but also subscriber dynamic data like the value of some counters at a certain point in time, his/her life-cycle state, or the status of his/her valid tariff options.

Such an approach supports scenarios like the following: as long as the subscriber consumption within one month does not exceed a certain limit, he/she is eligible for a better Quality of Service (QoS) than once the threshold has been exceeded. Alternatively, a scenario might occur in which a specific subscriber bought on top of his/her standard tariff an option for data traffic so that he/she is eligible for a better policy than "normal" subscribers.

Consequently, the SPR would have to store such information as well. However, this information is still mandatory in the OCS because it might influence ratings. For example, the high value of a usage counter respectively having subscribed to a certain tariff option might lead to a reduced or negligible price for data traffic. Or taking the example above again, once the subscriber consumption within one month exceeds a certain limit (not necessarily the same limit as for policy decision), the subscriber might enjoy cheaper rates for data traffic.

This shows that some subscriber data is meaningful both for the SPR and the ABMF. There could be here a kind of overlapping between the SPR and the ABMF as represented in Figure 1.

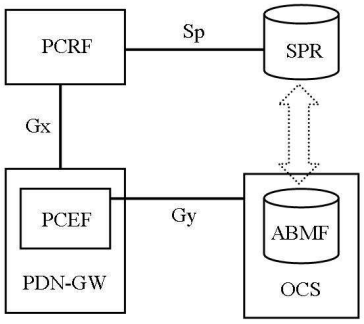


Figure 1. Potential overlapping between the SPR and the ABMF

Replicating the information both in the SPR and in the ABMF would be an option. But this would assume efficient synchronization mechanisms between the two databases, since the number of subscribers respectively their data traffic in today's telecommunication networks might be substantial. Furthermore, the involved pieces of information consist of real-time data. If the policy should change when the subscriber's consumption reaches a certain limit, the change would happen in real-time and without delay. In the same way, if the rating should change when a certain limit is reached, the change should happen in real-time too.

Duplication of databases, which store a great deal of real-time data, could increase the complexity of the implementation. If the relevant subscriber information is already present in the OCS, why should not the PCRF retrieve it directly from the OCS? This is represented in Figure 2.

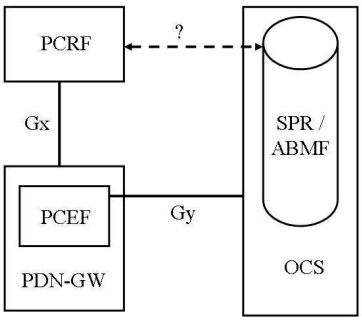


Figure 2. OCS acting as an SPR

III. PROPOSED APPROACH

The proposed approach consists of a framework where the PCRF and the OCS exchange in real-time subscriber information, which is necessary not only for charging, but also in order to determine the right policy. The goal is to support such scenarios where the policy might be changed based on the value of some subscriber data volume counters.

The latter are stored in the OCS as master copy in any case because they are relevant for charging, in order to support offers like the following: after a subscriber has consumed 1MB within one week, he/she gets 10 free SMS, or he/she is allowed free data traffic till the end of the week. Furthermore, these counters are relevant to the PCRF in order to support similar offers where, for example, the data speed is throttled once the subscriber has reached 10MB consumption within one month. In the context of the present contribution, we shall focus on volume counters. However, it could be another piece of subscriber data, which would be relevant for the policy server, for example, the life-cycle state of the subscriber. For example, if a prepaid data card is near expiry, the surfing speed may diminish.

In the context of the implementation described in the next section, these are the values of subscriber volume counters, which should be reported in real-time from the OCS to the PCRF. More precisely, the counter values will be reported when they exceed some predefined thresholds. The latter might be defined either for a certain subscriber marketing category, or for all the subscribers in the same tariff, or individually at subscriber level. Since these thresholds might be reached in the middle of a session, the OCS might have to notify the PCRF in the middle of a data session too.

Nevertheless, the PCRF should retrieve latest subscriber information like the tariff plan ID and the values of the

 ISBN:978-1-61208-113-7

volume counters at the beginning of the session as well in order to determine correctly the initial policy. Alternatively, the PCRF could replicate this subscriber information, meaning again that some synchronization mechanisms would have to be implemented.

In general, the message flow when a data session is established would resemble Figure 3. In (1), the PCEF asks the PCRF about the policy that should apply to the session, which is about to start for this subscriber. For this purpose, the PCRF retrieves latest subscriber information from the OCS in (2) and (3). Consequently, the PCRF can notify the initial policy to the PCEF in (4). This would happen through the Gx interface in accordance with [2].

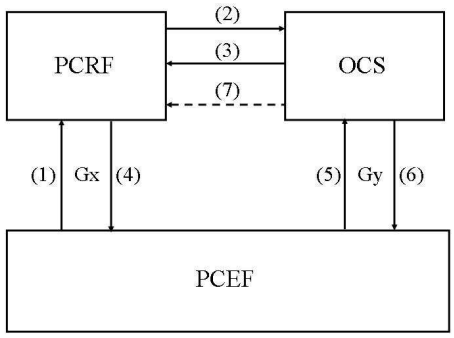


Figure 3. Message flow with PCRF/OCS interaction

Once the policy has been determined, the PCEF requests the OCS for a volume slice in (5). After checking the current subscriber consumption, the subscriber's default tariff respectively available options, and current balance, the OCS allocates a slice in (6). This would happen through the Gy interface in accordance with [2]. In order to allocate the proper slice, the OCS takes into account charging-relevant thresholds, but it should take into account policy-relevant thresholds as well: this will ensure a timely charging or policy change. Depending on the duration of the session, there might be several volume slices requested i.e., several messages like (5) and (6).

The arrow in (7) is represented in dotted line because it may or may not occur during a session: the OCS would notify the PCRF only if a policy-relevant threshold is exceeded during the on-going data session.

As stated above, the protocol for (1) & (4) respectively (5) & (6) is Gx respectively Gy. The protocol for (2) & (3) respectively (7) will be discussed in the next section. Since (2) & (3) respectively (7) are not fully covered by standard bodies yet to the best of our knowledge, the protocol which is the most convenient will be assessed.

An alternative approach, trying to stick to existing standards, would have been for the PCRF and the OCS to exchange information through the PCEF i.e., through the Gx and Gy protocols. But this would imply an extension of the existing protocols as well. In fact, since the present implementation, various alternatives are being discussed in [6]. They focus on the exchange of information about volume or monetary counters. Additional pieces of information such as the subscriber's life cycle state or his/her optional tariff options may become relevant too.

IV. IMPLEMENTATION

Regarding the protocol for (7) in Figure 3, since Gx and Gy rely on Diameter, and Gy on Diameter Credit Control Application [7], it was decided to use Diameter Credit Control Request (CCR) Event. The reader might have noted that in (5) & (6), the OCS acts as a Diameter Server towards its client i.e., the PCEF, while in (7) the OCS acts as a Diameter Client toward the Diameter Server, which is the PCRF in this case. As there might be several PCRF nodes, the OCS should support an N+K PCRF architecture in order to ensure a good scalability. The OCS should be able to send CCR Event messages to the PCRF nodes in round-robin way in order to ensure high-availability, meaning that the functionality can still be supported, even if one PCRF node is down.

Regarding (2) and (3), it is about the PCRF's retrieving subscriber profile data from the OCS database at the beginning of a session. Therefore, it is not really about Credit Control, nor Authentication / Accounting. Consequently, Diameter was not chosen, but SOAP/XML instead, because it is a simple protocol to let applications exchange information over HTTP in a platform-independent manner. For more information on SOAP/XML, the reader might refer to [8] and [9].

Within this framework, the following scenario can be supported: let us assume that a subscriber is entitled a downlink/uplink speed of 768/384 Kbps as long as he/she has not exceeded 10MB within a month. Once he/she reaches 10MB, he/she should be throttled to 128/64 Kbps. Let us assume that at the beginning of a session, the subscriber has a consumption of 9.9MB in the current month.

Consequently, when the session is established, the PCRF communicates a QoS corresponding to 768/384 Kbps to the PCEF. In addition, the OCS allocates a quota of only 0.1MB (10-9.9) in the initial Credit Control Answer (CCA) message. That way, when the threshold of 10MB is reached, the PCRF can be notified in real-time. This is represented in Figure 4.

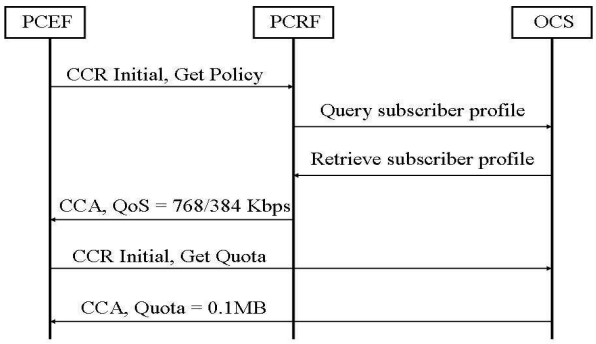


Figure 4. Initial slice granted by the OCS at session start

 ISBN:978-1-61208-113-7

In case the PCRF has a local database duplicating the OCS database, and containing subscriber information that is not outdated, the query from the PCRF to the OCS may be skipped.

When the allocated quota of 0.1MB has been used up, the PCEF should request another volume quota. If the subscriber balance is sufficient, the OCS will allocate another quota so that the data session can carry on. The allocated quota might be bigger than 0.1MB this time, for example 0.5MB.

Simultaneously, the OCS will notify through a Diameter CCR Event message as suggested previously that the volume threshold of 10MB has been reached for this subscriber, so that the PCRF can calculate the new QoS and notify it to the PCEF. This is represented in Figure 5.

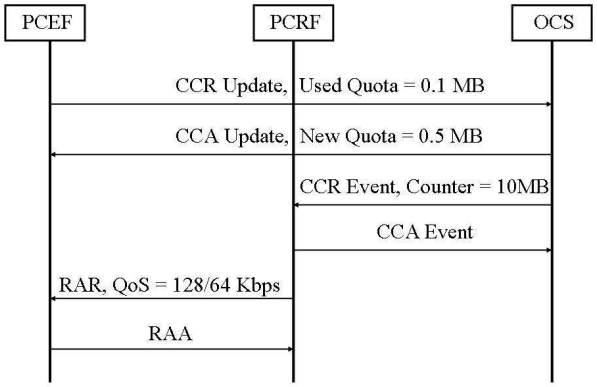


Figure 5. Mid-session notification from the OCS to the PCRF

In order to further notify the policy's change to the PCEF, the PCRF uses Diameter Re-Authentication Request / Answer messages (RAR/RAA) [10].

In case of multiple parallel sessions, the policy change should apply to all on-going sessions. For example, let us assume that one session – Session 1 – starts when the counter value is 9.9MB. Given the threshold of 10MB, the OCS should allocate initially a slice of 0.1MB. Before the latter is used up, another session – Session 2 – starts. The OCS also allocates 0.1MB as initial slice because the counter value is still 9.9MB in the OCS database. This is represented in Figure 6.

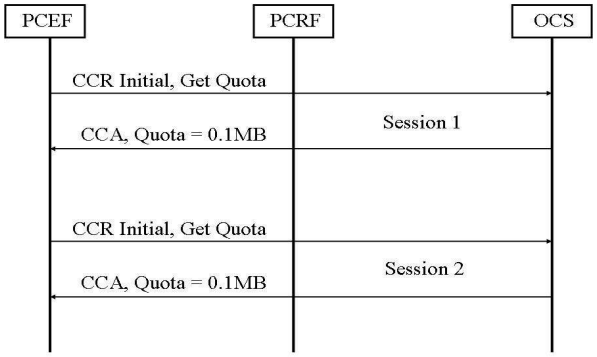


Figure 6. Initial slice for parallel sessions

As soon as the initial slice of 0.1MB of Session 1 or Session 2 is used up, the PCEF will request another slice. The OCS will grant a new slice, but it will update the volume counter value to 10MB, which should trigger the notification to the PCRF. This is represented in Figure 7, where the first session using up the 0.1MB quota is Session 1.

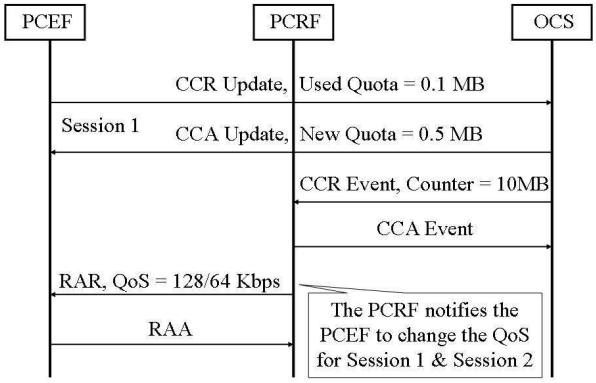


Figure 7. Mid-session QoS notification for parallel sessions

Consequently, the PCRF should notify the PCEF to change the QoS obviously for Session 1, but for Session 2 too, since the volume threshold is applicable to both Session 1 and Session 2, even if it was triggered by Session 1 only.

V. CONCLUSION

The 3GPP defines a valuable framework in order to grant different traffic policies applicable to different kinds of data traffic typologies. The policy server has the ability to retrieve subscriber information from a subscription repository in order to make individual policy decisions. However, the policy server could interact directly with an online charging system in order to support scenarios where the policy depends on subscriber real-time information, which is mandatory for rating and charging too. For this reason, we allowed ourselves to extend the 3GPP framework available at the time of the design, and implemented the described proposal.

In the IMS standard in Release 10, subscriber data is still present in different network elements depending on the application for which this data is required. When it comes to dynamic subscriber data required by different applications, dispatching might not always be suitable. Therefore, as in the approach presented in this paper, it will lead to the specification of new interfaces.

TERMINOLOGY

3GPP	3rd Generation Partnership Project
ABMF	Account and Balance Management Function
CCA	Credit Control Answer
CCR	Credit Control Request
GW	Gateway

 ISBN:978-1-61208-113-7

GGSN	GPRS Gateway Support Node
GPRS	General Packet Radio Service
GW	Gateway
Gx	IMS reference point between PCEF & PCRF
Gy	IMS reference point between PCEF & OCS
IMS	IP Multimedia Subsystem
IP	Internet Protocol
Kbps	kilo bit per Second
MB	Mega Byte
OCF	Online Charging Function
OCS	Online Charging System
PCC	Policy and Charging Control
PCEF	Policy and Control Enforcement Function
PCRF	Policy and Control Resource Function
PDN	Packet Data Network
QoS	Quality of Service
RAA	Re-Authentication Answer
RAR	Re-Authentication Request
RF	Rating Function
SOAP	Simple Object Access Protocol
Sp	IMS reference point between PCRF & SPR
SPR	Subscription Profile Repository
XML	eXtended Markup Language

ACKNOWLEDGMENTS

Much of the source material used for this paper derives from work accomplished together with the technical teams of Alcatel-Lucent, Openet and Vodafone. The author would like to thank especially Renée Fang, Xiang Yang Li, Jacob Terpstra, Simone Thomann, Gianmarco Montini from Alcatel-Lucent, Michael Leahy from Openet, Ron Jennissen, Jeffrey Lammerts van Bueren, Dion Pirnaij from Vodafone, Paul Grooten from EDS and Christian Falckenberg from Starent.

The author would like to thank Marianne Cave for her review.

REFERENCES

[1] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Online Charging System (OCS): Applications and Interfaces", 3GPP TS 32.296, Release 10, October 2010

[2] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Policy and Charging Control (PCC) architecture", 3GPP TS 23.203, Release 10, September 2010.

[3] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Policy and Charging over Gx Reference Point", 3GPP TS 29.212, Release 10, September 2010

[4] R. Good, and N. Ventura, "Application driven Policy Based Resource Management for IP multimedia subsystems", 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities (TridentCom), 2009

[5] T. Grgic, K. Ivesic, M. Grbac, and M. Matijasevic, "Policy-based Charging in IMS for Multimedia Services with Negotiable QoS Requirements", 10th International Conference on Telecommunications (ConTEL), 2009

[6] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "Study on Policy solutions and enhancements", 3GPP TS 23.813, Release 10, October 2010

[7] H. Hakala, L. Mattila, J-P. Koskinen, and M. Stura, J. Loughney, "Diameter Credit Control Application", IETF RFC 4006

[8] SOAP Version 1.2, World Wide Web Consortium (W3C) Recommendation, http://www.w3.org/TR/soap12-part1/ [retrieved: November 9th, 2010]

[9] World Wide Web Consortium (W3C), http://www.w3.org/standards/xml/ [retrieved: November 9th, 2010]

[10] P. Calhoun, G. Zorn, D. Spence, and D. Mitton, "Diameter Network Access Server Application", IETF RFC 4005

StepRoute - A MultiRoute Variant Based on Congestion Intervals

Ali Al-Shabibi
Kirchhoff-Institute for Physics
University of Heidelberg
Heidelberg, Germany
Email: ali.al-shabibi@cern.ch

Brian Martin
European Organization for Nuclear Research
Geneva, Switzerland
Email: brian.martin@cern.ch

***Abstract*—Congestion-aware routing protocols require network statistics which are timely and as precise as possible. Consequently, there is a need to represent congestion information efficiently and in a scalable manner. Based on our previous work, we propose a routing protocol, called StepRoute, which achieves these objectives, while retaining the functionality of routing according to local and remote network conditions. By classifying congestion values into different categories, we are able to deliver timely information to routers while retaining a meaningful estimate of the original statistical value. We compare our results with our previous work (MultiRoute) as well as shortest path only routing. We show that our new variant outperforms both shortest path and MultiRoute in terms of throughput. The protocol itself is media independent, but for test purposes we have employed Ethernet.**

***Keywords*—Communication systems, Protocols, Monitoring, Computer network performance, Communication system routing.**

I. Introduction & Previous Work

Redundant connections are common in modern networks. While these connections provide varying degrees of resiliency, they also deliver an opportunity for multipath protocols. Currently deployed routing protocols only consider shortest paths between any source-destination pair and seldom do they consider the congestion present within the network. Historically, attempts to deploy congestion-aware routing with the Arpanet [1] failed because of route flapping which would lead to out of order packets and therefore drastic performance degradation. Nevertheless, it has been shown that congestion control has a significant positive impact on routing performance [2].

Protocols, such as Open Shortest Path First (OSPF) [3], Routing Information Protocol (RIP) [4], or Interior Gateway Routing Protocol (IGRP) [5], all rely on the existence of a single path between any source-destination pair. Doing so causes them to always route packets for a given destination on the same path and thereby increase the build up of congestion. We consider networks where multiple paths exist, whether they are of equal length or within an acceptable margin of the shortest path. Traffic flows are then assigned to a particular path according to the congestion indication received from neighboring routers. In essence, we consider that both local and remote congestion information should be considered when performing a routing decision.

StepRoute is a multipath routing protocol which provides a lightweight mechanism to represent both local and remote congestion in a scalable and precise manner. The three components which make up StepRoute are:

1) **Path Construction:** By relying on the existence of a shortest path between any source and destination point, we have developed our multiple path discovery process. After establishing the shortest path cost (the reference cost), each alternate path is computed whose cost is within a reasonable delta of the reference cost. This ensures that the latency versus throughput trade off is respected. This process has been described in [6].
2) **In-Network Monitoring:** To ensure fresh and timely statistics the routers poll themselves, rather than having an external monitoring process poll them. The precision at which the congestion is represented is discussed in Section II-A, but we can safely say that our representation is significantly lighter than the one used in [7]. These statistics are then sent to neighboring routers via an aggregation protocol similar to [8], which enables the statistics to be distributed within the network efficiently.
3) **Routing Table Representation:** Each router is responsible for maintaining its own routing vector based on the congestion information of its local links and of neighboring routers. The congestion information sent, via the In-Network monitoring

 ISBN:978-1-61208-113-7

protocol, by neighboring routers must be interpreted correctly by the recipient router. We present a data structure, called a *routing mask,* which allows routers to interpret information correctly while enabling flexibility on the precision of the information it contains.

Based on the above components, routers initially discover the paths that are available to the different networks. Then, using the In-Network Monitoring protocol and the *routing masks*, routers are able to construct their routing table for each destination. It is important to notice that besides the path discovery the only factor in the routing decision is the congestion statistics, this enables the routers to update the routing tables as statistics become available. Therefore, we can precompute the routing tables which enables rapid next-hop lookups. Traffic is then grouped into flows which are identified by several parameters, and assigned to the port corresponding to their next-hop. The assignment of a flow to a port is immutable for the duration of that flow's lifetime. This simple approach avoids path oscillations and thus out of order packets.

Currently, routing protocols make inefficient or no use of congestion to route onto alternative paths. That said, there has been significant research in this field. Using Constrained Shortest Path First [9] over Multiprotocol Label Switching (MPLS) [10] is a traffic engineered [11] approach to load balanced routing, which requires the *a priori* knowledge of the traffic matrix. Our approach is generic and requires no *a priori* knowledge about the network topology nor traffic distribution. In [7], the authors propose a method which is similar to our approach but in which all the router-router link statistics for a given device are packed into a single value. We believe that such an approach causes a significant loss of precision and propose a method for providing statistics for all router-router links. In [12], the authors mainly address the problem of route oscillations and propose to route long-lived IP flows on different paths than short-lived ones, whereas we propose no such distinction.

The remainder of this paper is structured as follows, first we will give more detail about the components that make up StepRoute. Then, we will detail StepRoute's routing algorithm. Finally, we present StepRoute's results when compared to Shortest Path Routing and MultiRoute.

II. Component Description

Each of the components described above handles a different area of the overall routing protocol, some are performed at the initialization time while others run continuously.

A. In-Network Monitoring

When designing a congestion-aware routing protocol, it is crucial to deploy a mechanism which delivers statistics as quickly as possible. Clearly, using standard centralized monitoring tools like SNMP [13] or sFlow [14] would not be appropriate as the time to gather the statistics and redistribute them to the router would exceed any timing constraints.

In order to guarantee the freshness requirement expressed above, we have decided to implement our own monitoring protocol based on the idea presented in [15]. The main advantages of our approach are expressed below:

- **Distributed:** Each router polls itself locally and generates a value representing the current level of congestion.
- **Update on change:** Updates are only sent when a change in congestion occurs, similarly to [15], thereby reducing the overhead needed by the protocol.
- **No Flooding:** All updates are only sent to neighboring routers which do not forward them.

Relying on the property that routers update their interface counters frequently (based on our research, interface counters can be safely queried at one second intervals [16].), we compute the difference between two consecutive updates and use the following formula to derive a congestion value.

$$\gamma = \Phi \cdot \frac{\Delta}{\Gamma} \tag{1}$$

where Δ is the difference between two consecutive updates, Γ represents the capacity of the link and finally Φ is a constant, which represents the sensitivity of this measure; the larger it is, the more rapidly the congestion measure will increase.

Taking the result obtained from Equation 1, we classify the congestion value into a number of categories according to the degree of precision required (shown in Table I). This classification simplifies the route calculation process [17]. In particular, as we will see in Section III, it allows for a very simple routing algorithm.

Class 1	$0.0 < \gamma \leq \alpha(1)$
Class 2	$\alpha(1) < \gamma \leq \alpha(2)$
Class i	$\alpha(i-1) < \gamma \leq \alpha(i)$
Class P	$\alpha(P-1) < \gamma \leq \alpha(P) = 1.0$

TABLE I: Congestion Classification.

Classifying the congestion values allows us to simply represent the status of the network to neighboring

 ISBN:978-1-61208-113-7

routers. Consider the situation where we would want to have m classes of congestion where m can be expressed as 2^n, then we only n bits per router-router link are needed to represent these four possibilities. Therefore as the number of possible classes grows exponentially, the space required to represent them only grows linearly. This approach allows us to describe many different congestion levels while employing a lightweight method for describing them. Moreover, routers which receive this information do not need to know any extra information, such as link speed or duplex status, about the sending router.

A router then packs all the values corresponding to its router-router links into a data structure which will be represented in Section II-B.

B. Routing Table Representation

The statistics which are received by a router take the form of a sequence of bits. While this representation is extremely space efficient, it poses one major issue: how does a receiving router interpret this information? More precisely, each router may present multiple links to a destination and varying precision for any given router-router link, therefore a mechanism is needed to allow routers to accurately interpret this information.

We propose the notion of a *routing mask*, which relies on the ordering of the routing tables. The actual ordering relation can be arbitrary as long as all participating routers use the same one, for example our implementation orders entries in the routing table by destination network. Initially, each router sends its *routing mask* to all its neighbors and this is only done once unless there is a failure in which case the entire algorithm recomputes.

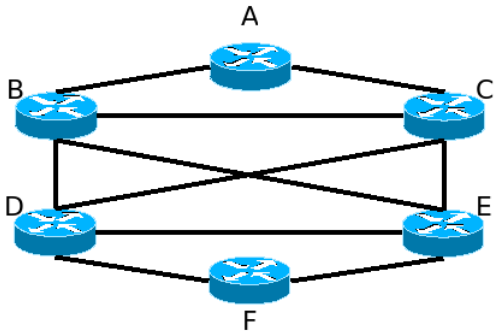


Fig. 1: The experimental network.

A *routing mask* consists of a sequence of zeros separated by ones. A consecutive sequence of zeros indicates remote links (plus the number of bits to represent the congestion class) which can be used to send packets to the same destination. A one indicates a transition to the next network in the routing table. The *routing mask* indicates to the receiving router the sequence of bits to expect for update messages received from remote routers.

		Destination					
		A	B	C	D	E	F
Source	A	L	1	1	2	2	2
	B	1	L	1	1	1	2
	C	1	1	L	1	1	2
	D	1	1	1	L	1	1
	E	2	1	1	1	L	1
	F	2	2	2	1	1	L

TABLE II: The connectivity table. Each entry shows the number of possible paths.

Considering the network given in Figure 1 and the connectivity map given in Table II, where each entry represents the number of paths to a destination network and L stands for locally connected, and finally assume that there are four classification classes and thus two bits are required to represent them. Under these condition router A would send the following routing mask to its neighbors **01001001000010000010000**, which indicates that A has two paths for networks D, E, and F. All updates send from routers will follow the same format, thereby enabling routers to immediately be able to interpret the incoming information and know exactly which part of the update vector is of interest to them.

III. Protocol Description

None of the components described above solve the main problem encountered by multipath protocols, namely, out of order packets which cause a drastic performance deterioration as is explained in [18]. We use the same approach that was used in MultiRoute [6] and in the OSPF variant Equal Cost MultiPath, where packets are classified into a flow by hashing the packet headers. We then use this classification to bind the flow to a given path. Once a flow is assigned to a path, it is bound to it for the duration of its lifetime and cannot be moved. Therefore, it is impossible to obtain out of order packets at the destination.

The three components described previously rely on each other to perform the objective function of the algorithm. The Routing Table representation depends on the In-Network monitoring protocol to deliver its message, and the monitoring protocol depends on the Path construction component to know to whom to send the statistics. All these components deliver their information to the routing algorithm, which takes routing decisions based on the available paths and their congestion status.

 ISBN:978-1-61208-113-7

StepRoute's routing algorithm requires the knowledge of the available paths (provided by the path construction component), local and remote statistics (provided by the In-Network Monitoring), and finally the *routing mask* (from the Routing Table Representation). The algorithm executes whenever new statistics are available, and re-populates the routing table for each destination. This pre-calculation enables rapid next hop lookup during the routers operation.

The routing algorithm consists of three cases corresponding to the state of the links local to a router:

- *Case I - All local paths uncongested.* In this case, the algorithm only looks at the statistics received from neighboring routers. Assuming there are multiple paths available, the router searches for the least congested path which is simple due to the classification of the congestion values discussed in Section II-A. Clearly, if the algorithm finds a remote path with is not at all congested, it immediately selects this path for forwarding. On the other hand, if all the remote congestion counts are equal or if none is found, the shortest path is selected.
- *Case II - Some local paths are congested while others are not.* This case is slightly more complex because a local path, even if it is carrying some traffic, may still be amongst one of the better options. This is due to the fact that remote paths, which lay beyond a completely uncongested link, may be completely congested. In this case, the algorithm ranks the candidate paths by summing their local congestion with the remote congestion. The path with the lowest congestion value is then selected. As with Case I, if there are multiple candidates, the shortest one is selected.
- *Case III - All local paths are completely congested.* This case is very much similar to the first case. The idea here is to look at the congestion values of remote routers and determine the least congested path, in an effort to use up all the available bandwidth. Again, if multiple candidates are found, the algorithm defaults to the shortest path.

Let us consider, as an example, the network given in Figure 1 and its associated Table II, when multiple flows enter at router F destined for network A. We also assume that each flow is long lived and that it immediately consumes half of the available bandwidth. There are four paths between networks F and A, namely F-D-B-A (1), F-D-C-A (2), F-E-C-A (3), and F-E-B-A (4), path (1) is considered to be the shortest followed by path (2) and so on. As there are only two distinct paths we can expect to double the network throughput with expect to a shortest path algorithm. It is important to note that, with respect to the real implementation the routing tables are pre-computed as statistics become available and not when a flow arrives.

When the first flow arrives, it is bound to path (1) as this path is considered to be the shortest, and due to Case I, this then causes an update from the routers F, D, and B indicating that their links are partially congested. When the next flow arrives, the decision is taken by Case II of the algorithm, and therefore the algorithm will consider the paths with next hop E as this link is not congested. The path (4) is excluded, because the link between B and A is congested due to the first flow. Therefore the second flow is bound to path (3). Upon arrival of the third flow, Case II will rank the available paths according to the congestion level and will choose path (2) as the link between D and C is not congested. Similarly, when the fourth flow arrives, Case II ranks the available paths again and picks path (4). As subsequent flows arrive at router F, Case III attempts to find available bandwidth to send the flow on and if this is not possible it sends it onto the shortest path.

IV. Results and Discussion

In this section, we present results obtained from our real world implementation. The experimental setup is shown in Figure 1 which operates at 100Mb/s and all link costs are equal to one. It was achieved using commodity routers running an OpenFlow [19] enabled firmware and using NOX [20] as the OpenFlow controller. The tests are performed by running 30-second UDP streams simultaneously between hosts connected to routers F and A using IPerf. The streams are run at 30 Mb/s and their number is increased to observe the behavior of StepRoute. We compare StepRoute (SR) (using four classes of congestion) with shortest path routing (SP), and MultiRoute(MR). Due to space requirements we are only able to present one set of results.

Figure 2 shows the performance of shortest path routing, MultiRoute and StepRoute under the scenario described above. Shortest Path routing is first to be limited, this is because for each flow SP routes onto the same path which means that by the third flow the path is nearly at saturation. MultiRoute performs better than SP, but still worse that SR. This is due to the fact that MR only uses a single bit to represent congestion and therefore a link may be marked as congested while it still has "spare" bandwidth and thus the protocol avoids using it. Finally, StepRoute does not suffer from this problem, because it classifies congestion into several classes, it has a finer control over the congestion levels of the different paths from F to A.

 ISBN:978-1-61208-113-7

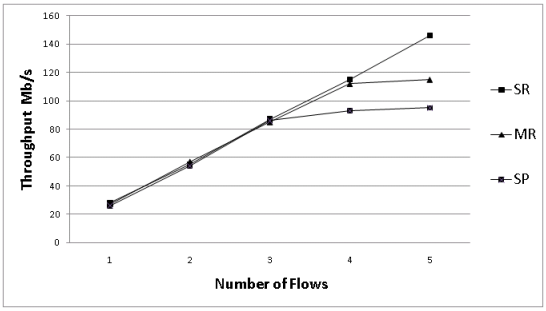


Fig. 2: StepRoute versus Shortest Path Routing and MultiRoute.

V. Conclusion and Future Work

In this paper, we have shown that by using congestion values, we can increase the overall throughput of a network. By using an in band monitoring protocol, we are able to deliver statistical information to routers in a timely manner. The use of classes of congestion greatly increases the performance of a protocol. More importantly, we are able to represent these classes efficiently with respect to a minimal one-bit marking. Our *routing mask* approach reduces to a minimum the amount of overhead required to signal router with detailed information about the network.

The results show that StepRoute performs significantly better than shortest path routing, and to our previous work; MultiRoute. Our results show that StepRoute is a promising protocol, but more work is needed to determine how well StepRoute scales when deployed onto a large-scale network. Also, it would be interesting to experiment with full-mesh traffic to observe StepRoute's behavior.

Future work is required to better understand the effects of the different parameters which make up StepRoute. Also, a future implementation of the *routing masks* would be to extend them to describe entire paths, from a source to a destination.

References

[1] A. Khanna and J. Zinky, "The revised arpanet routing metric," *SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 4, pp. 45–56, 1989.

[2] H. Rudin and H. Mueller, "Dynamic routing and flow control," *Communications, IEEE Transactions on*, vol. 28, no. 7, pp. 1030–1039, Jul 1980.

[3] J. Moy, "Ospf version 2," United States, 1998.

[4] C. L. Hedrick, "RFC 1058: Routing information protocol," Jun. 1988.

[5] B. Albrightson and J. Boyle, "Eigrp-a fast routing protocol based on distance vectors," in *Proc. Networld/Interop 94*, 1994.

[6] A. Al-Shabibi and B. Martin, "Multiroute - a congestion-aware multipath routing protocol," in *High Performance Switching and Routing (HPSR), 2010 International Conference on*, jun. 2010, pp. 88–93.

[7] I. Gojmerac, T. Ziegler, F. Ricciato, and P. Reichl, "Adaptive multipath routing for dynamic traffic engineering," *IEEE GLOBECOM*, vol. 6, pp. 3058 – 3062, Dec. 2003.

[8] R. Stadler, M. Dam, A. Gonzalez, and F. Wuhib, "Decentralized real-time monitoring of network-wide aggregates," in *LADIS*, New York, 2008, pp. 1–6.

[9] J.-L. L. Roux, J.-P. Vasseur, and J. Boyle, "Requirements for Inter-Area MPLS Traffic Engineering," RFC 4105 (Informational), Internet Engineering Task Force, June 2005. [Online]. Available: http://www.ietf.org/rfc/rfc4105.txt

[10] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol label switching architecture," United States, 2001.

[11] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "Overview and principles of internet traffic engineering," United States, 2002.

[12] A. Shaikh, J. Rexford, and K. G. Shin, "Load-sensitive routing of long-lived ip flows," in *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*. New York, NY, USA: ACM, 1999, pp. 215–226.

[13] J. D. Case, M. Fedor, M. L. Schoffstall, and J. Davin, "Simple network management protocol (snmp)," United States, 1990.

[14] M. Wang, B. Li, and Z. Li, "sflow: towards resource-efficient and agile service federation in service overlay networks," in *Distributed Computing Systems, 2004. Proceedings. 24th International Conference on*, 2004, pp. 628–635.

[15] A. Prieto and R. Stadler, "A-gap: An adaptive protocol for continuous network monitoring with accuracy objectives," *Network and Service Management, IEEE Transactions on*, vol. 4, no. 1, pp. 2–12, June 2007.

[16] S. Batraneanu, A. Al-Shabibi, M. Ciobotaru, M. Ivanovici, L. Leahu, B. Martin, and S. Stancu, "Operational model of the atlas tdaq network," *Nuclear Science, IEEE Transactions on*, vol. 55, no. 2, pp. 687 –694, april 2008.

[17] S. Bahk and M. El Zarki, "Dynamic multi-path routing and how it compares with other dynamic routing algorithms for high speed wide area network," *SIGCOMM Comput. Commun. Rev.*, vol. 22, no. 4, pp. 53–64, 1992.

[18] D. Thaler and C. Hopps, "Multipath issues in unicast and multicast next-hop selection," United States, 2000.

[19] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, 2008.

[20] N. Gude, T. Koponen, J. Pettit, B. Pfaff, M. Casado, N. McKeown, and S. Shenker, "Nox: towards an operating system for networks," *Computer Communication Review*, vol. 38, no. 3, pp. 105–110, 2008.

 ISBN:978-1-61208-113-7

A Naming Scheme for Identifiers in a Locator/Identifier-Split Internet Architecture

Christoph Spleiß, Gerald Kunzmann[1]
Technische Universität München
Department of Communication Engineering
Munich, Germany
{*christoph.spleiss, gerald.kunzmann*}*@tum.de*

***Abstract*—Many researchers agreed that splitting the IP-address into a locator and an identifier seems to be a promising approach for a Future Internet Architecture. Although this solution addresses the most critical issues of today's architecture, new challenges arise through the mapping system which is necessary to resolve identifiers into the corresponding locators. One interesting question is how the naming of identifiers is achieved. In this work we give an overview of a naming scheme for identifiers based on the HiiMap locator/ID split Internet architecture. The naming scheme supports user-friendly identifiers for hosts, content and persons and does not rely on DNS. We furthermore give a possible solution for a lookup algorithm that can deal with spelling mistakes and typing errors.**

***Keywords*-Locator/ID split; Future Internet; Naming schemes; Content Addressing**

I. Introduction

Today's Internet architecture has been developed over 40 years ago and its only purpose was to interconnect a few single nodes. No one expected that the Internet and the number of connected devices would grow to the current size. Measurements show that the Internet continues growing at a tremendous high rate [1]. The address space of the current IPv4 addresses is already too small to address every single node in the Internet and the growth of BGP routing tables sizes becomes critical for the Internet's scalability [2]. While IPv6 is a promising solution for the shortage of addresses, it will probably increase the BGP routing table problem. Besides that, more and more devices connected to the Internet are mobile, such as smart phones or netbooks. However, the current Internet architecture has only very weak support for mobility as the IP address changes whenever a device roams between different access points.

Separating the current IP address into two independent parts for reachability and identification is a possible solution to many problematic issues with today's Internet [3]. With this approach a known identifier (ID) can always be used to reach a specific host, no matter where it is currently attached to the network. However, not only the number of hosts has developed differently than initially expected, but also the way people use the Internet. Today, the focus of the Internet is on accessing a specific piece of information and the host that stores the information os of minor interest. Furthermore, the emergence of social networks, Web 2.0 applications, Voice over IP (VoIP) and instant messaging applications additionally put the person in the focus of interest.

The split of locator and ID thereby offers perfect prerequisites for the support of addressing schemes for content, information and persons. Using this paradigm, an ID is assigned for every host, content and person. A highly scalable and flexible mapping system translates IDs into the corresponding locators. Note that the mapping system is mandatory a part of each locator/ID split architecture.

A crucial question is how to name and assign IDs. As IDs are used as control information in communication protocols and packet headers, they are mostly fixed-length bit strings that can be hardly memorized by humans. In this work we present a flexible and adaptable naming scheme for IDs that can be used to identify hosts, content, persons and is open for future extensions. Our approach is based on the HiiMap Internet architecture [4]. HiiMap provides a highly scalable and customizable mapping system. It does not rely on the Domain Name System and allows each entity to calculate the requested ID on its own.

The paper is structured as follows. In Section 2 we discuss related work and different concepts of locator/ID split architectures. Section 3 describes our approach of a new naming scheme for IDs while Section 4 deals with a lookup algorithm that tolerates spelling mistakes. Section 5 summarizes the results.

II. Related Work

Many proposals dealing with the split of locator and ID have been published so far, but only a few of them discuss how to name IDs. However, almost all of them use a bit-representation of constant length as ID.

A. Host-based approaches

LISP: In contrast to other architectures that are examined in this work, LISP [5] does not separate the identifier from routing purposes. Within an edge network, the normal IP-address still serves as so called *Endpoint Identifier (EID)* and routing address at the same time. While the EID is only routable inside a LISP-domain, an additional set of addresses is used for the routing between different LISP-domains,

[1]G. Kunzmann is now working for DOCOMO Communications Laboratories Europe GmbH, Landsberger Strasse 312, Munich, Germany.

 ISBN:978-1-61208-113-7

which are called *Routing Locators (RLOC)*. RLOCs are the public IP-addresses of the border routers of a LISP-domain, globally routable, and independent of the nodes' IP addresses inside the domain. Whenever a packet is sent between different LISP-domains, the packet is first routed to the *Ingress Tunnel Router (ITR)*, encapsulated in a new IP packet, and routed to the *Egress Tunnel Router (ETR)* according to the RLOCs. The ETR unpacks the original packet and forwards it to the final destination. A mapping system is necessary to resolve foreign EIDs (EIDs that are not in the same domain) to the corresponding RLOCs. However, as in normal IP networks, the EID changes whenever a node changes its access to the network. Furthermore, DNS is still necessary to resolve human readable hostnames to EIDs.

HIP: The Host Identity Protocol [6] implements the locator/ID split by introducing an additional layer between the networking and the transport layer. For applications from higher layers the IP address is replaced by the *Host Identity Tag (HIT)*, which serves as identifier. The IP address is purely used as locator. The main focuses of HIP are security features. This is why the HIT is a hash value from the public key of an asymmetric cryptographic key pair. Encryption, authenticity and integrity can be achieved in this way. However, the coupling of ID and public key is a major drawback, as the ID changes whenever the key pair changes. Furthermore, HIP solely is a host based protocol and is not suitable for addressing content or persons.

HIMALIS: Like HIP, the HIMALIS (Heterogeneity Inclusion and Mobility Adaption through Locator ID Separation in New Generation Network) [7]approach realizes the locator/ID split by introducing an extra layer between network and transport layer, the so-called Identity Sublayer. HIMALIS can use any kind of addressing scheme for locators and supports security features based on asymmetric keys. However, it does not burden the ID with the semantic of the public key. HIMALIS uses domain names as well as host IDs to identify hosts. In contrast to other approaches, a scheme how to generate host IDs out of the domain name using a hash function is shown. However, they use multiple databases for resolving domain names and hostnames to IDs and locators. Furthermore, it is again only a host based protocol.

B. Content-based approaches

Contrary to the host based approaches, the **NetInf** (Network of Information) architecture shows how locator/ID separation can be used for content-centric networking [8]. By introducing an information model for any kind of content, NetInf allows fast retrieval of information in the desired representation. Thereby, each information object (IO) includes a detailed description of the content and its representations, with locators pointing to the machine that stores the information. The ID is assigned to the IO and is composed out of hash values of the content creator's public key and a label created by the owner. In order to find a specific IO, the creator's public key and label must be known exactly.

Another Future Internet Architecture focusing on content is **TRIAD** [9]. One key aspect of TRIAD is the explicit introduction of a content layer that supports content routing, caching and transformation. It uses character strings of variable length as content IDs and uses the packet address solely as locator.

C. Hybrid approaches

A proposal for a Next Generation Internet architecture that supports basically any kind of addressing scheme is the **HiiMap** architecture [4]. Due to the locator/ID separation and a highly flexible mapping system, HiiMap allows for addressing hosts as well as content and is still open for future extensions and requirements. In the following we use the term *entity* for any addressable item.

The HiiMap architecture uses never changing IDs, so called UIDs (unique ID) and two-tier locators. One part of the locator is the LTA (local temporary address) that is assigned by a provider and routable inside the provider's own network. The other part is the gUID (gateway UID). This is a global routable address of the provider's border gateway router and specifies an entrance point into the network.

HiiMap splits the mapping system into different regions, whereby each region is its own independent mapping system that is responsible for the UID/locator mappings of entities registered in this region. The mapping system in each region consists of a one-hop distributed hash table (DHT) to reduce lookup times. As DHTs can be easily extended by adding more hosts, the mapping system is highly scalable. In order to query for UIDs which regions are not known, a region prefix (RP) to any UID is introduced. This RP can be queried at the so-called Global Authority (GA), which resolves UIDs to RPs. The GA is a centralized instance and acts as root of a public key infrastructure, thus providing a complete security infrastructure. As RP-changes are expected to be rare, they can be cached locally.

Like other approaches, HiiMap uses fixed length bit strings of 128 bits as UID. As plaintext strings are not feasible as UIDs due to their variable length, a naming scheme is necessary to assign UIDs to all kinds of entities. Thereby, the existing Domain Name System is to be replaced by the more flexible HiiMap mapping system.

III. New naming scheme for Identifiers

In this section we introduce a naming scheme for IDs that is suitable to address basically any entity and that can be generated out of human friendly information. Although we use the HiiMap architecture exemplarily, this approach can also be adapted to other locator/ID split architectures.

 ISBN:978-1-61208-113-7

Type	input to Hash(name)	Ext 1	Ext 2
static host	plain text domain name	hash of local hostname	service
non-static host	global prefix assigned by provider	hash of local hostname	service
content	plain text content name	child content	version number
person	first + last name	random	communication channel

Table I: Content of UID fields corresponding to different types

A. General Requirements for Identifiers

When introducing a Future Internet Architecture based on locator/ID separation, the ID has to fulfill some mandatory demands. In the following we sum up general requirements for IDs proposed by the ITU [10]:

- The ID's namespace must completely decouple the network layer from the higher layers.
- The ID uniquely identifiers the endpoint of a communication session from anything above the transport layer.
- The ID can be associated with more than one locator and must not change whenever any locator changes.
- A communication session is linked to the ID and must not disconnect when the locator changes.

In addition to the ITU we add further requirements:

- An ID must be able to address any kind of entity, not only physical hosts.
- Every communication peer can generate the ID of its communication partner out of a human readable and memorable string.
- The ID is globally unique, but it must be possible to issue temporary IDs.
- The registration process for new IDs must be easy.
- IDs must be suitable for DHT storage.

While some of these aspects mainly affect the design of a Future Internet Architecture based on a locator/ID split, some issues are directly related with the naming of IDs.

B. Generalized Identifier

As IDs are used in the transport layer protocol to determine the endpoint of a communication, we cannot avoid using fixed-length bit strings to realize packet headers of constant size. In combination with DHTs, which also require fixed-length bit strings, the usage of a hashing function is obvious. In contrast to other approaches, which compose the ID of one hash value only, we split the ID in several predetermined fields whose purposes are known to all entities.

In the following we introduce a generalized scheme how to compose global unique IDs (UID) for any entity and give concrete examples how to name hosts, content and persons. Our scheme allows storing all these IDs in the same mapping database and is yet flexible enough to support different databases for different types of IDs.

Figure 1 shows the generalized structure of an ID, which is composed of a region prefix (RP) and an UID. The UID consists of a type field (T), the hash value of a human friendly name for the entity to be identified as well as two extension fields (Ext 1 and Ext 2). The UID is stored in the mapping system of a specific region, denoted by the RP.

The type field T denotes to which type of entity the UID belongs to. As T contains the most significant bits (MSB) in the UID, it is possible to map different ID types to different databases in the mapping system. We suggest 128 bits for the UID, whereby 4 bits are used to determine the type, 76 bits are assigned for the hash value, 32 bits for Ext 1 and 16 bits for Ext 2. In the following we show realizations for applying UIDs to different types: host, content and persons. Table I gives an overview how the UID is composed according to the type of entity. Each part is described in detail in the following subsections. Note that our scheme is not limited to these types, but can easily be extended.

Figure 1: UID with regional prefix RP

C. Identifiers for Hosts

IDs for hosts are the most common use case today and DNS is used to resolve hostnames to IP addresses in order to access a specific machine. The hostname, or FQDN (full qualified domain name), which specifies the exact position in the tree hierarchy of the DNS, can be roughly compared to the ID in a locator/ID separated Internet architecture. However, the FQDN is not present in any lower layer network protocol and is solely used in the application layer.

Similar to today's hostnames, we introduce a hierarchy to our UIDs. However, contrary to FQDNs, our scheme is limited to two hierarchical levels: a global part and a local part. While the global part is used to identify a whole domain, e.g. a company or an institute at a university, the local part is used to identify single machines within this domain. Note that the term *domain* does not refer to a domain like in today's DNS hierarchy. A domain in our solution has a flat hierarchy and simply defines an authority for one or more hosts. We differentiate between two different types of host UIDs:

1) Static Host Identifier: Static host UIDs are never changing IDs that can be generated by hashing a human readable hostname. Their main purpose is for companies or private persons that want to have a registered UID that is always assigned to their host or hosts.

Hash: The domain name part of the plain text hostname is used to generate the hash field of the UID.

 ISBN:978-1-61208-113-7

Ext 1: The hash value of the local host name is used to generate this field. A local hostname is unique inside a specific domain. An example for a local hostname could be *pc-work-003* and *mydomain.org* as its domain name.

Ext 2: The Ext 2 field is used to identify a specific service on the host. It can be compared to today's TCP or UDP ports. However, specifying a value in Ext 2 is not necessary when requesting the locator for a specific host from the mapping system and can therefore be set to zero. As the host is precisely identified with the global and local UID part, it is not necessary to store identifiers for each service of the host as they would all point to the same locator. Instead, Ext 2 is set to zero in the UID when querying the mapping system and filled with the specific service identifier when actually accessing the node.

For privacy reasons it is possible not to publish the UID for a private host in the global mapping system but only in a local database. For a single point of contact it is possible to use an UID with Ext 1 set to zero, which points to e.g. a login server, router or load balancer which forwards incoming requests to internal hosts.

Note that the host has to update its mapping entry pointing to new locator(s) upon a locator change.

2) Non-static Host Identifier: Contrary to static host IDs and the basic idea of never changing UIDs there will always be the need for non-static host UIDs, i.e. IDs that do not have to be registered, that are assigned to a host for a specific time and that are returned to the issuer if no longer needed. An example can be a private household with a DSL or dial-up Internet connection and a few hosts connected through a router. Each host needs its own, distinct UID to make connections with other hosts in the Internet, but it does not need to have a registered, never changing UID if no permanent accessibility is needed.

Hash: The global part is assigned to the router or middlebox that provides Internet access to the other hosts during the login process. It can be compared to the advertisement of an IPv6 prefix. The global part is valid as long as the customer has a contract with its provider. A new global part is assigned if the customer changes its provider. Yet, the transfer of a global UID part between different providers should be possible. In order to assign non-static UIDs to customers, each provider holds a pool of global UID parts. The mapping entry for a specific dynamic host UID is generated by the corresponding host immediately after assignment and whenever its locator changes. However, each host with no static UID assigned must proactively request a non-static host UID, either by its provider or router and middlebox, respectively. Note that the global part of a non-static host UID does not consist out of the hash value of a plaintext string and can therefore not be computed.

Ext 1: The local part of the dynamic UID is generated from the local hostname of a machine.

Ext 2: Identical to the static host UIDs.

D. Identifiers for Content

As the focus of the users in the Internet is shifting from accessing specific nodes to accessing information and content, different approaches towards a content-centric network have been made as shown in Chapter II. By applying the idea of information models, like the NetInf approach, to our naming scheme, each content, which can be e.g. a webpage or an audio or video file, gets its own distinct UID. Hereby, the UID does not point to the data object itself but to the information model of the content that has a further description and metadata stored.

Hash: For generating the UID of content we have to use a meaningful name that can describe the corresponding content or information. While this is indeed quite a difficult task, possible solutions could be e.g. the name of a well-known newspaper like *nytimes* which refers to the front page of the New York Times online version. Similar, the name of an artist could refer to an information object where albums or movies are linked.

As the spelling of the content description is not always exactly known, we suggest a lookup mechanism that can cope with minor spelling mistakes in the next chapter. In our proposal this plaintext name is used as input to a known hash function to generate the hash part of the UID.

Ext 1: This field is optional and can be used to access some more specific parts of the content or information that is directly related with the main object. This can be e.g. one specific article from a newspaper site or one specific album or piece of music from an artist. Ext 1 can help to avoid downloading a maybe bigger object description of the main content to gather the desired information. Another benefit is that each child object has its own locator and therefore can be stored on different locations while still being accessible through its parent UID. This is not possible today as e.g. the URL of a newspaper article is directly coupled with the host storing the information.

Ext 2: This field can be used to access a specific version of the desired content or information. Like in a versioning system, the Ext 2 field allows the user to easily access any earlier version and the changes made to the information. The actual version can be obtained by setting Ext 2 to zero.

Unlike with host addressing, we cannot simply connect to a locator returned by the mapping system. As the information object is a description of content or information, the requesting application or user has to evaluate the information object and select the desired representation according to the users needs. Thus, the network stack will not evaluate the data received from the mapping system for a content UID query but forward it to the corresponding application.

Note, n case a name, e.g. *nytimes*, refers to both a host (company) as well as content (webpage), the type field is used to differentiate whether a host or a content locator is returned.

 ISBN:978-1-61208-113-7

E. Identifiers for Persons

With the emergence of social networks, Internet-capable devices, Voice-over-IP (VoIP), etc., the need for personal IDs arose, as the *person* inself is moving in the focus of interest. Whenever somebody wants to contact a specific person he is interested in the communication with that person and does not want to care about the device, e.g. which phone or computer the person is currently using for communication. However, the user must have the possibility to choose the communication channel. That can be an email, a phone call, a message on a mailbox, a chat with an instant messenger or a message in a social network and so on.

Hash: The main part of a person's UID consists of a hash value calculated from the person's full name, i.e. first name plus last name. As many people have the same first and last name, the hash value is ambiguous and we need further information to distinguish between different persons.

Ext 1: For this purpose we use a random number for Ext 1 when initially generating a person's UID [11]. This initial generation is not done by the person itself, but is issued by a federal authority and valid for lifetime.

Ext 2: This field is used to specify the communication channel to the corresponding person and has a set of predetermined values, e.g. for email, VoIP, or instant messaging. Note, there are still enough unused values for future needs. According to each Ext 2 value, different locators can be stored in the mapping system, i.e., the Ext 2 value referring to the VoIP account can point to the locator of a VoIP provider or directly to a VoIP phone, the value referring to the mailbox can point to a mail server. The mapping entry for Ext 2 set to zero includes the person's full name and, depending on the person's privacy settings, further details about the person like birth date or current residential address. Ext 2 set to one is used to get the locator of the machine the person is currently working on if the corresponding person agreed to publish this information. Thereby, the communication channel can be signaled in a higher layer.

However, to contact a specific person, not only the person's name but also Ext 1 must be known. There are two possibilities: First, the initiating person knows the correct UID of its communication partner because they have exchanged it (like email addresses today). Second, the holder of a personal UID can agree to be indexed in a directory that is accessible through a personal UID with Ext 1 and Ext 2 set to zero. This directory can be compared to a phone book and stores additional information about all persons that have the same name including their random Ext 1 values.

IV. Identifier assignment and lookup

As each UID is globally unique by definition, it must be ensured that only one entity at a time has a specific UID assigned. It must be further prevented that any entity is hijacking an UID for malicious purposes.

A. Registration and assignment

The registration process for a static host UID can be compared to today's domain names. Whenever a new static host UID shall be registered, the corresponding mapping region checks, like the NIC today, if a specific UID is already registered. If the UID is unused, the mapping region creates an initial entry in the mapping system, including the host's public key. From now on, the host can update its mapping entry at any time, e.g. when it changes its access point. The update message to the mapping system must be signed with the host's private key, thus avoiding the UID to be hijacked. The UID at the node is configured via a system file like /etc/hostname. The owner of a host must proclaim changing the key-pair of a node at the mapping region.

The purpose of non-static UIDs is that they do not need a registration process, as their prefixes are assigned by a provider and therefore belong to that provider. However, it must be possible for hosts with non-static UIDs to change their mapping entries due to roaming although they have not been individually registered. Therefore, whenever a new non-static host UID is assigned, the provider creates the initial mapping entry on behalf of the corresponding host using the host's public key. Then, the host can directly update its locator at any time in the mapping system. However, if a host wants to change its key pair, the node must directly be connected to the provider. Only the provider can verify that this host is allowed to update the public key because the provider can verify the login data. If a host is permanently relocated to another provider, it has to request a new non-static host UID at its new access point. After that it has to initiate the clearance of its old UID.

The procedure for content UIDs is basically the same like for static host UIDs. The content creator has to initially register the hash part of the UID at the mapping system. However, it does not need to register each single content that is provided. Then the content provider can freely create new content that only differs in Ext 1 and Ext 2. A special case is content that is free to public changes like Wikipedia. Here, everybody is allowed to create a new version of the corresponding content that differs in Ext 2 but changes must be verified with the person's key pair.

Unlike with UIDs for hosts or content, UIDs for persons are assigned by an authority of the state. As the personal UID can be used to make transactions and legal contracts, it has to be guaranteed that the UID cannot be abused. Furthermore, it has to be guaranteed that values for Ext 1 are unambiguous. That would not be the case if everybody would generate its own random value for Ext 1. Thus, during the registration process, an authority creates the mapping entry for the person requesting a UID and deposits the person's public key. Then, the person can update and create any entry for Ext 2 on its own. Changing the key pair must be accomplished through the issuing authority.

 ISBN:978-1-61208-113-7

B. Lookup mechanism

The idea of our naming scheme for IDs is based on the fact that each UID can be generated out of a known plain text string with a known hash function and without an additional naming system like DNS. However, as the main part of any UID consists of a hash function, the desired entity can only be found if the plain text string that builds the UID is exactly known and no spelling mistake or typing error occurred. To overcome this drawback, we suggest a lookup mechanism that is based on n-grams in addition to the pure UID lookup.

1) n-gram generation: Although DHTs only support exact-match lookups, it is possible to use n-grams to perform substring and similarity searches. Hereby, each plaintext string is split up into substrings of length *n*, which are called *n-grams*. The n-gram and its hash value is then stored as key/value pair in the DHT [12].

A typical value for n is two or three. With $n = 3$, the content name `nytimes` e.g. is split up into $I = 5$ trigrams h_i with $i = 1, ..., I$: `nyt`, `yti`, `tim`, `ime`, `mes`. Additional to the actual mapping entry indexed by the UID, the hash value $H(h_i)$ of each n-gram h_i is inserted in the mapping system together with the corresponding plain text name P. Thereby, the mapping entry for an n-gram consists of the tuple $\langle H(h_i); plaintext\ string \rangle$ [11]. Although these tuples are stored in the same mapping system like the UID, we suggest using a different database within the mapping system for performance reasons. Whenever the entity changes its location, no updates of the n-grams are necessary, as they do not contain any locator information but only the entitie's plaintext name.

2) Querying UIDs: Whenever querying the mapping system for a specific UID, the first step in the lookup process is using the precalculated (or already known) UID as query parameter. Only if the mapping system is not able to find a mapping entry to the corresponding UID, e.g. because of a spelling mistake, the n-gram lookup is executed. It is up to the user or application if an n-gram based query request is initiated.

In doing so, the second step is to calculate the corresponding n-grams out of the plaintext string and query the mapping system for each n-gram. The mapping system sorts all matching n-grams according to the frequency of the plaintext string and returns the list to the user. With high probability, the desired plaintext has a high rank in the returned list. By further correlating the input string with each returned plaintext string, the result is even more precise [13].

As the user must evaluate the results returned by an n-gram query, the network stack will forward that data directly to the application, which is responsible for correct representation. However, although this feature is similar to Google's "Did you mean...?", the mechanism is not suitable to handle complex queries with semantically coherent terms as Google can do.

V. Conclusion

In this work we presented a new naming scheme for IDs in locator/ID separated Future Internet Architectures. The generalized ID scheme is suitable for basically addressing any kind of entity. We showed examples for hosts, content and persons. Because each UID can be computed out of a human readable plaintext string, an additional naming system like DNS is not necessary any more. Due to the extendible type field, we have the possibility to assign ID-types for e.g. mobile phones, sensors or even cars or abstract services that provide any functionality to a user. Because IDs are independent from locators, a communication session is not interrupted upon an access point change. Furthermore, by introducing an n-gram based extended lookup mechanism we are able to cope with spelling errors and typing mistakes, thus improving the quality of experience for the user.

References

[1] ISC, "The ISC Domain Survey," http://www.isc.org/solutions/survey, Internet System Consortium, 2010.

[2] A. Afanasyev, N. Tilley, B. Longstaff, and L. Zhang, "BGP routing table: Trends and challenges," in *Proc. of the 12th Youth Technological Conference High Technologies and Intellectual Systems¨*, Moscow, Russia, April 2010.

[3] B. Quoitin, L. Iannone, C. de Launois, and O. Bonaventure, "Evaluating the benefits of the locator/identifier separation," in *Proc. of 2nd ACM/IEEE Internat. Workshop on mobility in the evolving Internet architecture*. ACM, 2007, pp. 1–6.

[4] O. Hanka, G. Kunzmann, C. Spleiss, and J. Eberspächer, "HiiMap: Hierarchical Internet Mapping Architecture," in *1st Internat. Conf. on Future Information Networks*, 2009.

[5] D. Farinacci, V. Fuller, D. Oran, D. Meyer, and S. Brim, "Locator/ID separation protocol (LISP)," Draft, 2010. [Online]. Available: http://tools.ietf.org/html/draft-ietf-lisp-07

[6] R. Moskowitz and P. Nikander, "Host identity protocol (HIP) architecture," RFC 4423, Tech. Rep., May 2006.

[7] V. Kafle and M. Inoue, "HIMALIS: Heterogeneity Inclusion and Mobility Adaptation through Locator ID Separation in New Generation Network," *IEICE TRANSACTIONS on Communications*, vol. 93, no. 3, pp. 478–489, 2010.

[8] C. Dannewitz, "NetInf: An Information-Centric Design for the Future Internet," *Proc. 3rd GI/ITG KuVS Workshop on The Future Internet*, May 2009.

[9] D. Cheriton and M. Gritter, "TRIAD: A new next-generation Internet architecture," Tech. Rep., 2000. [Online]. Available: http://www.dsg.stanford.edu/triad

[10] ITU, *Draft Recommendation ITU-T Y.2015: General requirements for ID/locator separation in NGN*, 2009.

[11] G. Kunzmann, "Performance Analysis and Optimized Operation of Structured Overlay Networks," Dissertation, Technische Universität München, 2009.

[12] M. Harren, J. Hellerstein, R. Huebsch, B. Loo, S. Shenker, and I. Stoica, "Complex queries in DHT-based peer-to-peer networks," *Peer-to-Peer Systems*, pp. 242–250, 2002.

[13] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *6th European Conf. on Speech Communication and Technology*, 1999.

 ISBN:978-1-61208-113-7

Address-Translation-Based Network Virtualization

Yasusi Kanada, Toshiaki Tarui

Central Research Laboratory, Hitachi, Ltd.
Higashi-Koigakubo 1-280, Kokubunji, Tokyo 185-8601, Japan
{Yasusi.Kanada.yq, Toshiaki.Tarui.my}@hitachi.com

***Abstract* – Two network-virtualization architectures, namely, network segmentation and network paging, were investigated. They are analogical to two memory-virtualization architectures: segmentation and paging. Network paging, which is relatively new and is based on a type of network-address translation (NAT), is focused on. This architecture requires smaller packet size and has several more advantages over the conventional architecture (i.e., network segmentation). Intranet- and extranet-type communication methods based on this architecture are described. An address translators is placed at each edge router in the WAN and used to evaluate client-server communication under wide-area virtual-machine (VM) live migration as a case of extranet-type communication.**

***Keywords* – network virtualization; segmentation; paging; network address translation; NAT; extranet.**

I. INTRODUCTION

Network virtualization (NV) isolates multiple communities while using the same hardware, namely, computers, network nodes, and network links. It enables users to create their own wide-area networks. Virtual networks (VNs) are customizable and programmable. Because the developers of VNs can exclude the complicated and unnecessary features of conventional internet-protocol (IP)-based networks, the structure of a VN is much simpler. The developers can use simplified IP protocols such as IP−− [Oht 10] or can introduce non-IP protocols that are simpler, more powerful, and more efficient.

One of the complicated functions performed by real-world IP-based networks is network-address translation (NAT) [Zha 08] [Ege 94]. Using NAT is limited and complicated, so many engineers and scientists would prefer to avoid using it. However, it plays an important role in real-world networks. Conventional NAT [Sri 01] is useful when the number of available IP addresses is less than required. It is also useful when there are IP addresses that are only used locally or should be hidden from the global network.

Several types of address translation will play important roles in NV. Although all types of address translation can be called NAT, in this paper, the term "NAT" is not used for these types of address translation because NAT is usually used for conventional specific types of address translation, so it may cause misunderstanding. However, even in the case of conventional NAT, the localization and information hiding described above can be regarded as a virtualization function. Similar to dynamic address translation (DAT) used in memory virtualization, address translation is one of the two core functions that can be used for virtualization.

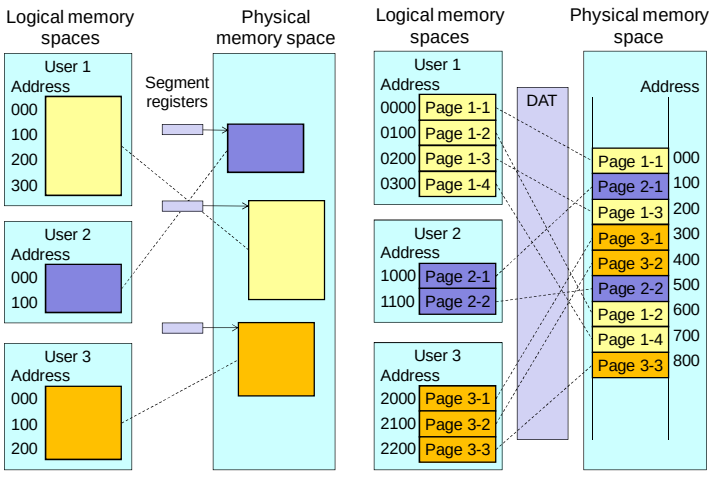
In the remainder of this paper, paging and segmentation in main-memory virtualization is explained and two NV architectures, namely, network-paging-based or address-translation-based architecture and network-segmentation-based architecture, are described in Section II. The former architecture is explained in detail in Section III, and a communication method using this architecture is described in Section IV. An application of address-translation-based virtualization, namely, wide-area VM live-migration, is presented in Section V. Related work is briefly reviewed in Section VI, and the paper is summarized in Section VII.

II. PAGING AND SEGMENTATION

A. Paging and segmentation in main memory

Virtualization technology was first developed for virtualizing computer memory. In particular, data in memory was read and written by using virtual addresses. Two memory-virtualization architectures, segmentation and paging [Tan 08], were developed.

- *Segmentation*: A memory-virtualization architecture in which the memory space is divided into logically separated and variable-sized segments and each user uses a segment (see **Figure 1**(a)). Logical and physical memories are mapped to each other by using segment registers that point to the head of physical-memory segments. A memory address is represented by a pair consisting of a segment (register) number and a displacement in the segment.
- *Paging*: A memory-virtualization architecture in which the memory space is divided into fixed-size pages and the pages of all the users of a computer are mapped into a sin-



(a) Segmentation (b) Paging

Figure 1. Two memory-virtualization architectures

 ISBN:978-1-61208-113-7

gle large address-space (see Figure 1(b)). Logical and physical memories are mapped to each other by using dynamic address translation (DAT). A memory address is represented by a number that indicates a point in the address space.

These architectures have advantages and disadvantages: segmentation is conceptually simpler, but paging is simpler to implement. They may also be used in combination.

B. Paging and segmentation in network

We can assume that there are two NV architectures that correspond to the above two memory-virtualization architectures because of the analogy between memory- and network-virtualization described below. In the case of memory virtualization, memory data are organized into a virtual-memory structure that differs from a real-memory structure. Similarly, in the case of NV, objects such as virtual machines (VMs) are organized into a virtual-network structure that differs from a real-network structure. Multiple memory-address spaces are created by memory virtualization. Similarly, multiple network-address spaces (or name spaces), in which virtual hosts, virtual nodes, and other virtual objects are identified, are created by NV. Both memory data and packets are read and written by using virtual addresses (or names), and the formats of data addresses and object addresses are similar too. The following two NV architectures are therefore assumed.

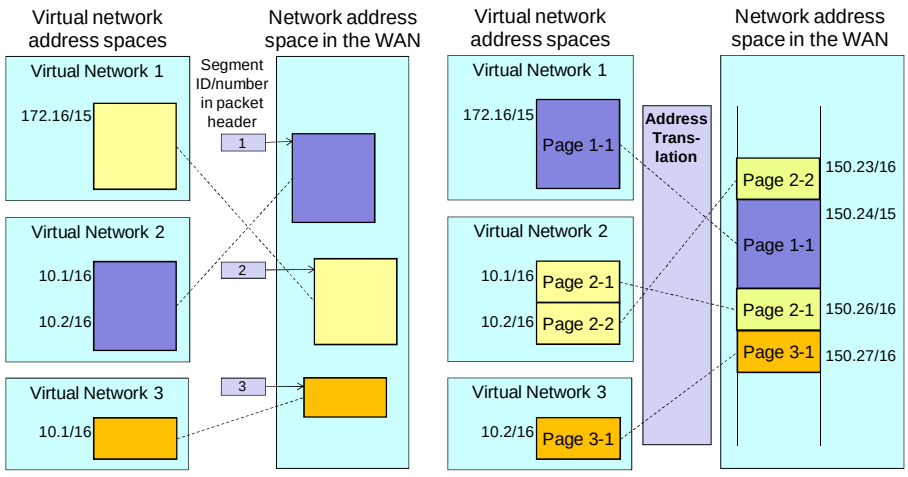
- *Network segmentation*: A NV architecture that distinguishes every network object by a pair, namely, a VN identifier (called a *segment identifier*) and a virtual address (or name), called an object *identifier* (OID) hereafter. VPN numbers or names, or VLAN identifiers, are used as segment identifiers (see Figure 2(a)). A real-network address is represented by a pair of these two identifiers, and each packet contains the sender's and the receiver's OIDs of this type. If identical OIDs are used in two VNs, they can be distinguished because their segment identifiers are different. This type of virtualization is widely used in VPNs and experimental virtual networks.
- *Network paging* (*address-translation-based virtualization*, *ATV*): A VN architecture that distinguishes every network object in all VNs by a single unique address. The OIDs are mapped into the address space of a wide-area (or global) network (WAN). This mapping is a type of NAT. A real-network address is represented by this single address, and each packet contains the sender's and the receiver's OIDs of this type. A virtual-address space may be divided into multiple pages and may be mapped to two or more non-contiguous subspaces in the WAN (see **Figure 2**(b)), although the page size may be varied because there is no hardware restriction. Each VN page must be mapped to a non-overlapping range of the WAN address space. If the same OIDs are used in two VNs, they are mapped to different addresses in the WAN.

Most conventional NV methods are based on network segmentation. Each data frame in a VN is encapsulated by a packet header of the substrate network (i.e., underlying network), and the segment identifier is in the packet header. Typical NV methods use IP-based encapsulation such as generic routing encapsulation (GRE) [Far 00], use layer-2 methods such as VLAN, or use multi-protocol label switching (MPLS). With these methods, GRE keys, VLAN tags, or MPLS labels contain the segment identifiers or labels that correspond to the segment identifiers.

In contrast to segmentation, network paging (i.e., ATV) seems to have been seldom used for communication between two or more sites of a VN. Conventionally, each local network behind a NAT is an independent network site, so it is not regarded as a VN site.

The format for identifiers is the same as that for network segmentation and network paging; that is, the addresses (or names) can be structured as a pair, i.e., (PS, F). PS represents the segment or page, and F represents the sub-address or field identifier that distinguishes the object from other objects in the same page or segment. Both PS and F may be numbers or symbolic names such as a fully qualified domain name (FQDN). Examples of PS and F are given as follows.

- *Numerical example*: PS = 172.16/16 (the first 16 bits of an IPv4 address) and F = *.*.10.21 (the last 16 bits). PS represents a network page.
- *Symbolic example*: PS = example.com and F = www. PS represents a segment.
- *Compound example*: PS = Government (a segment name) and F = 02.01.043 (a sub-address that consists of a department, a division, and a host number).



(a) Network segmentation (b) Network paging

Figure 2. Two network-virtualization (NV) architectures

Three differences between network segmentation and network paging are explained here. The first is the overhead caused by segmentation and paging. In segmentation, a segment identifier is added or removed at LAN-WAN borders. The packet size becomes larger in the

 ISBN:978-1-61208-113-7

WAN; on the other hand, it is smaller in paging (i.e., LAN). However, the processing overhead of adding or removing a segment identifier is not large. In contrast, the processing overhead of address translation may be large.

The second difference concerns the size and number of segments and pages. In the case of paging, the WAN address space can be divided into a collection of pages, where the size and number of pages can be altered. However, in the case of segmentation, the size and number of segments cannot be changed because each segment is a logically separated address-space.

The third difference is between the methods for handling non-IP protocols. In regard to network segmentation, the encapsulated payload may contain a data frame of an arbitrary format. It may contain a non-IP frame; namely, non-IP protocols can be handled. In regard to network paging, non-IP protocols can also be handled, but the translators must map non-IP addresses to IP addresses if the WAN uses IPs.

Network segmentation and network paging can be combined. For example, paging can be used for data packets, and segmentation can be used for packets of the ICMP and those of routing protocols.

III. ADDRESS-TRANSLATION-BASED VIRTUALIZATION (ATV)

A. Requirements of ATV

Two conditions are required to enable ATV.

- *Identity of addresses in VN sites*: Addresses (OIDs) used for the same object at each site of a VN must be identical in ordinary cases. This means the mapping at the exit of the WAN must be the inverse of that at the entrance.
- *Isolation of VNs*: The address translator at the entrance of the WAN may not let a packet with a disallowed address pass through. Multiple addresses used for a VN must be mapped into different addresses in the WAN. In addition, the address translator at the exit of the WAN may not let a packet with a disallowed address pass through. These conditions are met by dropping a packet when no translation rule matches either the source or destination address.

An important difference between memory paging and network paging is that, in the latter, the address (OID) in an incoming packet must be translated in the inverse direction. Such inverse translation is not required in the former because incoming data, namely, memory content, do not contain an address.

B. Ordered/unordered addresses

With memory virtualization, the addresses are ordered. With NV, however, the OIDs are not necessarily ordered; that is, there are at least two types of identifiers.

- *Ordered identifiers*: The relationship between two identifiers, i1 and i2, is given as i1 > i2, i1 = i2, or i1 < i2. IP addresses are ordered identifiers because address ranges, or subnets, are meaningful.
- *Unordered identifiers*: No order between two identifiers is defined. The MAC addresses are unordered identifiers.

In the case of ordered identifiers, addresses on a page (in a certain range) can be translated using the same method as in memory virtualization. For example, because IP addresses are ordered, ATV using IP-to-IP translation is very similar to memory paging. The virtual address space can also be divided into multiple pages and mapped to non-contiguous addresses (see Figure 2(b)). In the case of IP addresses, a subnet can naturally be regarded as a page.

In contrast, unordered identifiers may have to be handled in a different way; that is, an output address may have to be specified for each input address. In such a case, the translation-table size must equal the number of MAC addresses in the virtual space. For example, MAC addresses are represented by a 48-bit number, but the order is not significant. Each MAC address may therefore have to be handled separately.

C. Types and varieties of mapping

Address translation is not necessarily restricted to the page-to-page type; that is, a translator may map contiguous addresses to non-contiguous addresses. If the virtual space is symbolic, the mapping is also non-numerical. However, numerical mapping is focused on here, and it is categorized into three groups. It is assumed that the original address is *ai* and the converted address is *ao*.

- *Contiguous translation*: This type of translation maps contiguous addresses to contiguous addresses. For example, $ao = ai + 100$. Figure 2(b) depicts this type.
- *Striped translation*: This type of translation maps contiguous addresses to striped addresses with constant strides. For example, $ao = 3 * ai + 1$.
- *Randomized translation*: This type of translation maps contiguous addresses to randomized addresses. For example, *ao* can be generated by a pseudo-random-number generator or a mapping table.

Randomized translation may be useful for security purposes because it makes address scanning difficult. Striped translation may be useful when there is a need to assign two or more WAN addresses to each virtual address or vice versa.

Three miscellaneous issues concerning mapping are described below. First, if the address format used in the VN is structured, it is possible to map part of the address that is functional in the WAN to the WAN address, and to store the rest, or whole address, in the payload. For example, if an address consists of a locator and a host-identifier, and the WAN is an IP network, the former can be mapped to an IP address, and the latter can be stored in the payload.

Second, an address in the VN can be translated into combination of addresses in two or more layers in the WAN; namely, the WAN address may contain information of individual hosts (i.e., MAC addresses). This representation probably works well in small-scale WANs such as enterprise-wide networks, but it is difficult to use them in a large-scale network because this representation is not scalable.

Third, in network paging, if the same type of address space is used for the VN and the WAN, and there is no address conflict, address translation is unnecessary (but an access control may be required) between them. An example of this case is given in Section V; however, this is a special case, and address translation is usually required.

 ISBN:978-1-61208-113-7

D. Advantages and disadvantages of ATV

The advantages and disadvantages of ATV compared to segmentation-based virtualization are listed here. The advantages are as follows.

- *No overhead and less redundancy in packets*: There is no overhead in terms of packet size and less redundancy in the WAN. For example, in the case of IP-to-IP translation, the packet is the normal IPv4 packet. In contrast, in the network-segmentation-based method, the packet must have a tunnel header, which contains the segment identifier, in the WAN.
- *Availability of WAN functions*: Virtualized packets may utilize WAN functions because the behavior of the packets depends on the WAN addresses; e.g., if the WAN is an IP network, the functions of ICMP or routing may be useful.
- *Availability of NAT implementations*: Although conventional NAT and address translation required for virtualization are different, implementations of the former may be enhanced to include functions required for the latter. In particular, because of the IPv4 address exhaustion problem, a high-performance carrier-grade (large-scale) NAT [Nis 09] will be deployed. It may be used for virtualization, and it will enable wire-rate translation performance.

The disadvantages of ATV are listed as follows.

- *Potentially large memory size and slow rate of processing*: Address translation requires rule memory (or translation-table memory) and long? processing time. The required memory size may be large. In that case, it may be difficult to process address translation at the wire rate.
- *Restriction on OID formats*: The OIDs of hosts or nodes in the VN must be mapped to addresses in the WAN. This may restrict the syntax and/or semantics of the OIDs.
- *Possible conflict with WAN function*: VN functions may cause conflict with WAN functions; e.g., if the WAN is an IP network, address translation may make routing work in an unexpected way on the VN.

Although the segmentation-based method seems to be easier and simpler in many cases, ATV has several advantages such as smaller packet size, flexible page size, and page-by-page processing. Several examples of network paging are presented in the next section.

IV. COMMUNICATION THROUGH A WAN

Two types of communication method between two VN sites through a WAN are shown in this sections. The first type involves intranet-type communication (i.e., two sites are in the same VN) and the second type involves extranet-type communication, where the two sites are in different VNs, but they are allowed to communicate with each other.

A. Intranet-type communication

A paging-based intranet-type communication (i.e., communication in a virtually closed network) is illustrated here. It is assumed that there is a VN with at least two sites (see **Figure 3**), which are connected through a WAN. It is also assumed that IPv4 is used in the WAN, but other protocols such as IPv6 or Ethernet (i.e., VLAN) can also be used. The sites can thus be connected through the WAN using IPv4.

If the WAN is a closed network, it is possible to put a translator at every external interface of the WAN edge routers to inhibit any unauthorized access to the VN. This set up is similar to a memory-paging architecture, because access control is a function of DAT, and usually no memory access can bypass the DAT. In contrast, if the WAN is an open network, such as the Internet, it is difficult to exclude unauthorized accesses. Because host addresses are mapped to the WAN address space, unauthorized users at a third site may illegally access the hosts. This can result in a security risk, so such access must be inhibited.

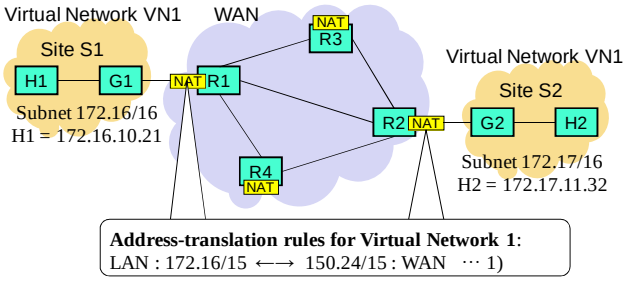


Figure 3. Intranet-type communication between two ATV-based VN sites

Figure 3 shows a translation rule that is required for communication between two sites, S1 and S2, through a VN. The same rule is used for both sites. It is applied to both the source and the destination addresses of a packet.

Only one page (a single rule) is used in this example. However, there may be multiple pages; for example, sites S1 and S2 may use different pages that map to non-contiguous pages in the WAN. If there are two or more rules, they are conceptually applied sequentially; namely, the first rule that matches the packet is used. However, the data structure of the rules can be optimized, and indexing or hashing may be used.

The meaning of the rule is described as follows.

- If a packet comes from the LAN, and if the source or destination subnet (the first 16 bits of the address) is 172.16/16 (or 172.17/16), it is translated to 150.24/16 (or 172.25/16); i.e., the rule is applied from left to right. The sub-address (the last 16 bits) is invariant (see Figure 2(b)).
- If a packet comes from the WAN, and if the source or destination subnet is 150.24/16, it is translated to 172.16/16; i.e., the same rule is applied inversely (from right to left), as described in Section III *A*. The sub-address is invariant.

Since there are no other rules, a packet with a source or destination address that is not specified in any rule is dropped.

Data packets are processed as follows. When host H1 at site S1 sends a packet to host H2 at site S2, the source address (172.16.10.21 in the figure) is translated into a WAN address (150.24.10.21) at ingress edge-router R1. The destination address (172.17.11.32) is also translated into a WAN address (150.25.11.32). This means Twice NAT [Sri 99] is applied to the packet. A Twice NAT is a type of NAT that modifies both the source address and the destination address. These addresses are IP addresses if the VN uses IP, but they may be another type of identifier if it uses a non-IP protocol.

 ISBN:978-1-61208-113-7

The simple contiguous translation described above was used. However, as described in Section III *C*, the address in VN1 can be separated into a subnet and a sub-address (host address), which are handled separately. Namely, the subnet can be mapped into a WAN address, and the sub-address can be put in the payload. This type of address translation is more like the conventional NAT that distinguishes local addresses by port numbers.

If dynamic routing is used in the VN, routing-protocol messages should be passed through the LAN-WAN borders. The messages will probably work well if dynamic routing is also used in the WAN. In this case, the edge routers of the WAN must translate the subnets in the messages when they pass through the address translation. When importing routes from the WAN to a VN site, they must be properly filtered. No routing-protocol extension is usually required to convey routes in the VN.

B. Extranet-type communication

A paging-based typical extranet-type communication (i.e., between intranets with access control) is illustrated here. There are assumed to be three VNs, i.e., VN1, 2, and 3 (see **Figure 4**). Each VN has only one site. Hosts at site S1 can communicate with hosts at the other two sites, S3 and S4. Hosts at site S3 can communicate only with hosts at S1 and S3, and hosts at site S4 can communicate only with hosts at S1 and S4. Figure 4 shows the translation rules, 1, 2, and 3 (three pages), required for communication between the three sites, S1, S3 and S4.

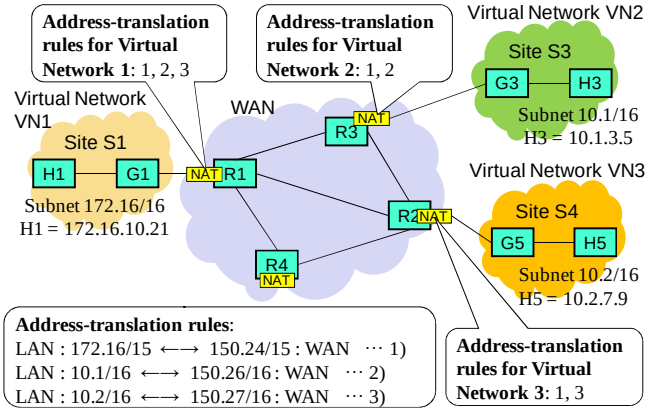


Figure 4. Extranet-type communication between two ATV-based VN sites

Hosts at S1 can communicate with hosts at S3 across the WAN. Rules 1 and 2 are used at edge routers R1 and R3 for this purpose. Data packets are processed using the following method (see Figure 2(b).) When host H1 at site S1 sends a packet to host H3 at site S3, the source address of the packet (172.16.10.21) is translated by using rule 1 for VN1 into a WAN address (150.24.10.21) at ingress edge-router R1. This translation is the same as in the intranet case. However, the destination address (10.1.3.5) is translated by using another rule, rule 2 for VN2, into a WAN address (150.26.3.5). The source- and destination-addresses are translated in the reverse direction at egress edge-router R3, so host H3 sees the original addresses. Hosts at S1 can also communicate with hosts at S4. Rule 1 and rule 3 for VN3 are used at edge routers R1 and R2.

On the contrary, hosts at S3 and S4 cannot communicate with each other because there are no rules for their communication in edge-routers R3 and R2. Namely, R3 does not have rule 3, and R2 does not have rule 2. For example, if a host address at S3 (10.1.3.5) is specified in a packet that comes from S4, the translator in the ingress edge-router R2 drops this packet because R2 does not have a rule that matches this address.

This type of access control can be specified page by page. This means that if a virtual-address space is divided into multiple pages, access to each page from outside the site can be controlled by the existence or non-existence of a rule for the page because each rule is defined for one page.

If there is no need to control access, rules 2 and 3 can be replaced by a single rule with doubled address ranges, which is similar to the rule used in the intranet example. However, they are separated for the purpose of access-control. If the WAN is the Internet, hosts in the VNs can communicate with hosts in the Internet if each router holds rules that map the hosts' addresses.

V. Application To VM Migration

To improve the performance of client-server communication under wide-area live migration of server virtual machines (VMs) (i.e., when server VMs are migrating between data centers [Kan 11]), the extranet-type communication method was applied. This method is briefly outlined in the following.

Wide-area VM live-migration between data centers can solve problems such as load balancing, disaster avoidance and recovery, and power saving. However, to enable migration between distant locations, other problems must be solved. One problem is "address warping". When a server VM is moved from one location to a more distant location, the IP and MAC addresses "warp" from the source server to the destination server. This confuses or complicates the status of both the WAN and LANs in a short time. This problem may cause a serious failure in the case of real-time traffic such as that involved in conferencing or on-line games.

This problem is solved by putting two data centers in different VNs, VN1 and VN3, as shown in **Figure 5**. This set up allows the VMs before and after migration that have identical IP and MAC addresses to briefly coexist. The IP addresses of the VMs are mapped to different addresses in the WAN, so no confusion occurs when the VM moves.

Users of the VM belong to another VN (VN2). When the VM is in source data center DC1, the address of the VM (172.16.10.21) in the users' VN is mapped to the VM before the migration from DC1. However, when the VM moves to destination data center DC2, a translation rule at the edge routers connected to the users' sites, R2 and R3, is switched, and the address of the VM in the users' VN is mapped to the VM after the motion in DC2.

The feasibility of ATV was tested by using IP-based simulated WAN and VNs. The WAN consisted of three layer-3 (L3) switches, translators (Linux PCs) connected to the L3

 ISBN:978-1-61208-113-7

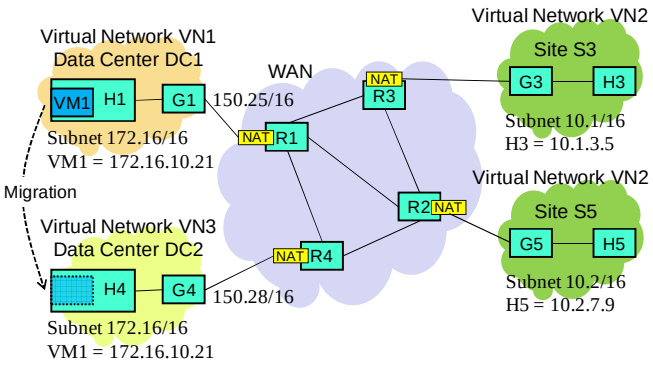


Figure 5. Wide-area VM migration using ATV-based VNs

switches, two sets of servers managed by VMware® at two simulated data centers, and a client PC at a simulated user site. The page size was 2^{16}. After the VM motion, RARP (reverse address resolution protocol) messages were generated by VMware and detected. The VN was then switched by a tool developed by an author [Kan 11].

In the current version of the translator, addresses in ICMP packets are not fully translated, so the function of the VNs is still limited. However, a UDP packet generator was embedded in a VM, and packets were captured at user sites. As a result, it was confirmed that a VM motion switched the VN and that the VNs worked correctly without any confusion.

VI. Related Work

NAT has been used to connect to the Internet by using only one IPv4 address or fewer addresses than the number of hosts. Because detailed standards are lacking, NAT has been developed in an ad hoc way. Much work has therefore been devoted to finding systematic methods for solving this problem. Among this work are IPNL [Fra 01] and IP4+4 [Tur 02], which enable communication between hosts behind the NATs. However, they were not intended to be used with virtualization applications.

Okada, et al. [Oka 02] described a method of deploying extranets using Twice NAT. This method is similar to our method described in Section V, but their major purpose was to deploy extranets without using multiple global addresses. They did not generalize their method for virtualizing networks.

Hasenstein [Has 97] mentioned the roles of NAT in NV and explained that he wanted to show where NAT might find or had already found its place in the entire virtualization scheme. Hara, et al. [Har 03] briefly mentioned the use of NATs in extranets. However, they did not discuss NAT as a virtualization medium.

VII. Conclusion

Two network virtualization (NV) architectures, namely, network paging and network segmentation, were described and compared. An address-translation-based virtualization (ATV, i.e., network-paging-based) method was investigated. Intranet- and extranet-type communication methods based on this architecture were proposed. An address translator was placed at each edge router in the WAN, and extranet-type communication during wide-area live migration of VMs was evaluated as a special case of extranet-type communication.

Segmentation has been widely used for NV, and segmentation-based methods would seem to be easier or more efficient in many cases. However, ATV-based methods have several advantages, such as less packet overhead, flexible page size, and page-by-page processing. As in the case of memory virtualization, segmentation and paging may also be used in combination to combine advantages of both architectures. Network paging is therefore a promising NV architecture. The authors will continue to develop and evaluate ATV-based methods, ways of utilizing the underlying network function in VMs, and a method for developing programmable networks based on ATV.

Acknowledgment

Part of the research results described in this paper is an outcome of the Eco-Internet Project (“R & D on Power-Saving Communication Technology − Realization of Eco-Internet −”), performed in fiscal year 2009, which was funded by the Ministry of Internal Affairs and Communications of the Japanese Government.

References

[Ege 94] Egevang, E. and Francis, P., “The IP Network Address Translator (NAT)”, RFC 1631, IETF, 1994.

[Far 00] Farinacci, D., Li, T., Hanks, S., Meyer, D., and Traina, P., “Generic Routing Encapsulation (GRE)”, RFC 2784, March 2000.

[Fra 01] Francis, P. and Gummadi, R., “IPNL: A NAT-Extended Internet Architecture”, *ACM SIGCOMM 2001*, pp. 69–80, August 2001.

[Har 03] Hara, Y., Ohsaki, H., Imase, M., Tajima, Y., Maruyoshi, M., Murayama, J., and Matsuda, K., “VPN Architecture Enabling Users to be Associated with Multiple VPNs”, *5th Asia-Pacific Symp. on Information and Telecomm. Tech.* (*APSITT 2003*), November 2003.

[Has 97] Hasenstein, M., “Diplomarbeit – IP Address Translation”, http://www.hasenstein.com/HyperNews/get/linux-ip-nat.html .

[Kan 11] Kanada, Y. and Tarui, T., “A “Network-Paging” Method for Wide-Area Live-Migration of VMs”, *25th International Conference on Information Networking* (*ICOIN 2011*), January 2011.

[Nis 09] Nishitani, T., Yamagata, I., Miyakawa, S., Nakagawa, A., and Ashida, H., “Common Functions of IP Address Sharing Schemes”, draft-nishitani-cgn-05, Internet Draft, IETF, July 2010.

[Oht 10] Ohta, M. and Fujikawa, K., “IP− −: A Reduced Internet Protocol for Optical Packet Networking”, *IEICE Transactions on Communications*, E93.B, No. 3, pp. 466-469, 2010.

[Oka 02] Okada, K., Chen, E. Y., Komiya, T., and Fuji, H., “Deploying User-based Extranet without Global Addresses”, *IPSJ SIG Notes*, CSEC 2002(43), pp. 7–12, Information Processing Society of Japan, May 2002.

[Sri 99] Srisuresh, P. and Holdrege, M., “IP Network Address Translator (NAT) Terminology and Considerations”, RFC 2663, IETF, August 1999.

[Sri 01] Srisuresh, P. and Egevang, K., “Traditional IP Network Address Translator (Traditional NAT)”, RFC 3022, IETF, January 2001.

[Tan 08] Tanenbaum, A. S., “Modern Operating Systems”, Third Edition, Pearson Prentice Hall, 2008.

[Tur 02] Turányi, Z. and Valkó, A., “IPv4+4”, *10th IEEE Int’l Conference on Network Protocols* (*ICNP’02*), November 2002.

[Zha 08] Zhang, L., “A Retrospective View of Network Address Translation”, *IEEE Network*, Vol. 22, No. 5, pp. 8–12, September/October 2008.

 ISBN:978-1-61208-113-7

Next Generation Access Networks (NGANs) and the geographical segmentation of markets

João Paulo Ribeiro Pereira
Polytechnic Institute of Bragança (IPB)
Bragança, Portugal
jprp@ipb.pt

Pedro Ferreira
Technical University of Lisbon (IST)
Lisbon, Portugal
pedrof@cmu.edu

***Abstract*—Telecom infrastructures are facing unprecedented challenges with increasing demands on network capacity. Next Generation Networks (NGN) allows consumers to choose between different access network technologies to access their service environment. The arrival of NGAN (Next Generation Access Network) has implications for the competitive conditions in access markets that are still uncertain (for example: access to ducts, dark fiber, equipment, etc.). The definition of the access price is a critical question, particularly when the incumbent also has activity in the retail market. In some regions, the regulatory authorities need to define the max price for wholesale access. In this context, the paper is divided into two main parts: 1) First we make a review of the main broadband access technologies (NGANs), and we propose a techno-economic model to support the new requirements of fixed and nomadic users. 2) In the 2nd part we propose a tool, developed in c language, which simulates the impact of retail and wholesale services prices variation in the provider's profit, consumer surplus, welfare, etc.**

Next Generation Networks, Next Generation Access Networks, Geographical Segmentation, Segmented Regulation, Nash Equilibrium

I. INTRODUCTION

The move towards Next Generation Networks (NGN) has begun to transform the telecommunication sector from distinct single service markets into converging markets [1]. Telecom infrastructures are facing unprecedented challenges, with increasing demands on network capacity. NGN allows consumers to choose between different access network technologies to access their service environment. In our work, the architecture of NGN will be limited to the current and future developments of network architectures in the access network (local loop), called Next Generation Access Network (NGAN). The NGAN can use technologies as fiber, copper utilizing digital subscriber line (xDSL) technologies, coaxial cable, powerline communications, wireless solutions or hybrid deployment of these technologies (Figure 1).

The choice of a specific technology for NGAN can be different between countries, geographic areas and operators. In recent years there has been an increase in the number, coverage and market share of "alternative" networks or operators such as resellers, unbundling operators, cable network operators, operators using frequencies for Wireless local loop (WLL), or operators deploying optical fiber in the local loop [2]. This has resulted in differences in competitive conditions between geographic areas, and this, in turn, has led to increasing argument (especially from incumbent operators) that geographical aspects be recognized in market/competition analysis and in regulatory decisions. There are several factors that can be responsible for this discrepancy [3]: state and age of the existing network infrastructure, length of local loop, population density and structure of the housing market, distribution of number of users and number of street cabinets for Local Exchange, level of intermodal competition in the market, willingness to pay for broadband services and the existence of ad hoc national government plans for broadband development.

The arrival of NGAN has implications for the competitive conditions in access markets that are still uncertain, including the role of bitstream, Sub-Loop Unbundling (SLU) access to ducts, etc. However, operator's investments in networks face different types of uncertainties. For example, when the incumbent operator has the monopoly in the access network and, simultaneous, has activity in retail market, the price regulation is an important question. Without price access regulation the incumbent can use his power in the market to stop or hamper the entrance of new operator in the retail market. However, if regulatory authority makes a very rigid control of the access price may reduce the incentive of the incumbent to make investments in the network. The regulatory authority should not increase uncertainties and has to provide clear incentives and guidance for the investment required for deploying NGANs [4]. Regulators should ensure that Local Loop Unbundling (LLU) and SLU, bitstream, the transition to NGAN, access to ducts and dark fiber, inside (building) wiring, collocation, and backhaul are defined in a transparent, efficient, and technologically neutral manner [2]. Segmented regulation has been identified as a regulatory framework that can potentially provide both incentives and controls for the deployment of NGNs [5].

Regulatory authorities in most OECD countries have traditionally adopted a national geographic area focus when framing the geographic scope of telecommunications markets [2]. The increase in the number, coverage and market share of new networks or operators has resulted in

 ISBN:978-1-61208-113-7

differences in competitive conditions between geographic areas. Results from market analysis economics suggest that differential regulation be considered between geographic areas where facility-based competition has developed and where it has not.

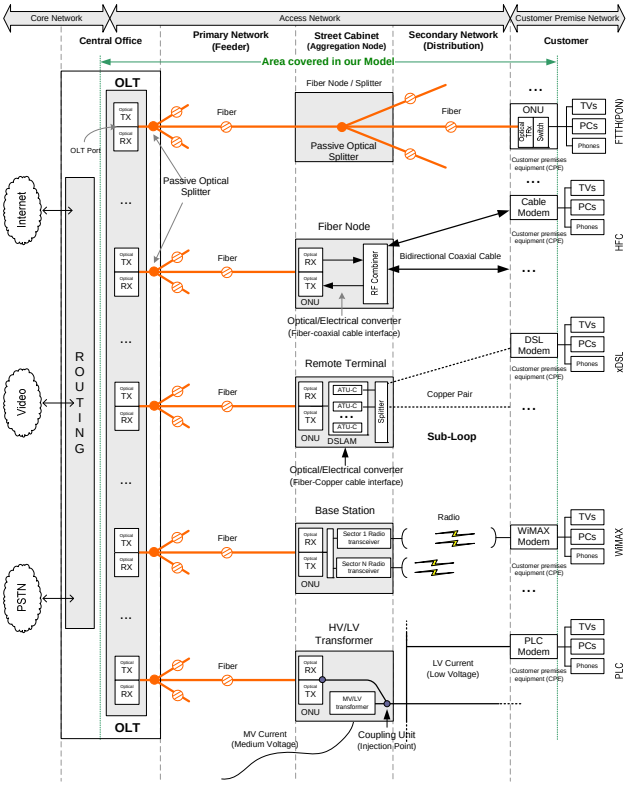


Figure 1. NGAN architectures (Block diagram) [6]

Competition can be promoted at many levels and locations through contestability and innovation [7]. After the decision of several countries to implement geographic regulation, the interest in these questions has been an increase. In the literature on the regulation of future access networks the discussion on regulation and investment has gained center stage given the pending infrastructure investments in many countries [1]. The geographically segmented regulation should aim not only at facilitating deregulation but also at strengthening regulation in those regions where competition is assessed to be ineffective. Then, segmented regulation can assist regulators to ensure that the regulatory framework they apply is appropriately tailored to the competition situation [2]. Local decisions of a national regulator may lead to inefficiencies deriving from discrepancies between local and global cost-benefit evolutions [8]. Segmented regulation may be helpful because it allows different solutions for the deployment of NGNs in urban and rural areas to evolve at different paces [5]

Figure 2 illustrates a scenario of the differences in competitive circumstances that may warrant geographically segmented regulation. There are geographical differences in conditions of competition: number of suppliers, market shares, etc. [9].

The deregulation of high-density areas may avoid unnecessary protection of access-based competitors and strengthen incentives to invest in infrastructure, and that maintaining regulation of low-density areas may promote competition with national offers, because alternative operators are enabled to extend geographical coverage

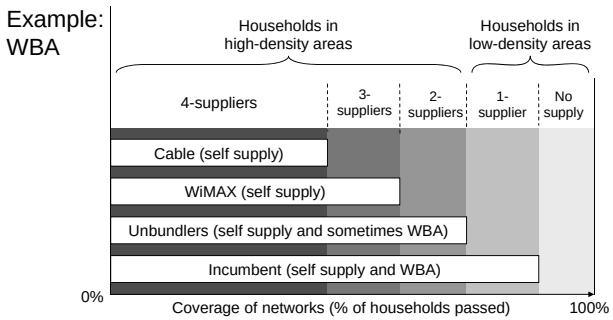


Figure 2. Geographically segmented regulation [9]

The analysis of some regulatory inquiries [1, 2, 7] on the national level shows that access providers (usually the incumbent operators – the former monopoly operators) are generally in favour of geographic differentiation. For example: the Spanish operator (Telefonica) claims that (see [2]) "…the geographical segmentation model will push investments and gradual deregulation and users will always enjoy the best possible scenario, either with a sustained or a regulatory supervised market…Differentiated regulation would prevent the increase of the digital divide." In Australia, Telstra argued that geographically segmented regulation (see [2]): "…will promote competition by giving service providers the appropriate incentives to use and extend alternative infrastructure, and will also promote competition in the upstream local services market by encouraging other carriers to offer wholesale local services."

In case of consumers, the geographic differentiation impact has an important consideration especially in view of the often-repeated statements by politicians and regulators that policy and regulation are designed to be in the long-term interest of consumers [1]. For business users, the breakup of market analysis to sub-national level is a source of significant alarm, especially concerning wholesale broadband access services. For multinational business users, inconsistency of national regulations, and a consequent inability to obtain seamless international network services without service quality, costs and administrative disadvantages, is already a serious problem.

II. A TECHNO-ECONOMIC MODEL FOR BROADBAND ACCESS TECHNOLOGIES

A. *Overview*

In this part we present the economics of next generation access networks, focusing on several broadband access technologies (fiber to the home - FTTH, DSL, hybrid fiber-cable - HFC, power line communication - PLC and worldwide interoperability for microwave access - WiMAX)

and propose a techno-economic model to support the new requirements of fixed and nomadic users.

We present an economic model designing and deploying access networks for both fixed and mobile users. The type of networks for fixed users includes, FTTH, DSL, HFC, PLC, while the nomadic user is assumed to use WiMAX. This model could serve as a good starting tool for the design of access networks, as it includes all the major capital expenses involved in the deployment of access networks such as equipment costs, installation costs, etc. The costs resulting for our model is grouped into 3 main categories: A) Infrastructures: in our work we subdivide this category in 2: civil works (trenching and ducting) and cable costs (cost of the fiber and cost to pass the cable in the ducts). As we can see in Figure 1, in all solutions the fiber costs are necessary to connect central office to the street cabinet. Normally, this is the category that has more costs, depending on the existence/share of an infrastructure. This means that if an operator has the infrastructure (or part), the costs can be lower. B) Equipment: includes the costs such as optical network unit (ONU), optical line termination (OLT) ports and chassis, splitters, digital subscriber line access multiplexer (DSLAM), etc. Also includes the cost of street cabinets and the equipment installation costs. C) Customer Premises Equipment: includes the modem and other electronics – like splitters. For FTTH architecture includes the ONU equipment.

B. *Description*

The proposed model considers that in the static layer, users are stationary and normally require data, voice, and video quality services (these subscribers demand great bandwidth). In the nomadic layer (or mobility layer), the main concern is mobility and normally the required bandwidth is smaller than in the static layer. The focus of the wireless networks was to support mobility and flexibility, while for the wired access networks is bandwidth and high QoS. However, with the advances in technology, wireless solutions such as WiMAX have capacity to provide wideband and high QoS services and in this way competing with wired technologies [10]. Then, we propose a new model to support the new needs of the access networks: bandwidth and mobility (see Figure 3).

For the nomadic layer we chose the WiMAX solutions. This technology enables long distance wireless connections with speeds up to 75 Mbps per second. WiMAX can be used for a number of applications, including "last mile" broadband connections, hotspot and cellular backhaul and high-speed enterprise connectivity for businesses. This technology can offer very high data rates and extended coverage.

As we can see in Figure 4, the framework is separated into three main layers [12]: (Layer 1) First, we identify for each sub-area the total households and SMEs (Static analysis), and total nomadic users (Mobility analysis).The proposed model initially separates these two components because they have different characteristics. (Layer 2) In this layer, it is analyzed the best technology, for each Access Network, the static and nomadic component. For the static analysis we consider the following technologies: FTTH-PON (passive optical networks), DSL, HFC, and WiMAX PLC. To the nomadic analysis we use the WiMAX technology. Then, the final result of this layer is the best technological solution to support the different needs (Static and nomadic). The selection of the best option is based in four output results: NPV, IRR, Cost per subscriber in year 1, and Cost per subscriber in year n. (Layer 3) The next step is to create a single infrastructure that supports the two components. To this end, is necessary the analysis of the best solution (based on NPV, IRR, etc.) for each Access Network. Then, for each sub-area we verify if the best solution is: a) the wired technologies (FTTH, DSL, HFC, and PLC) to support the static component and the WiMAX technology for mobility; or b) use the WiMAX technology to support the Fixed and Nomadic component.

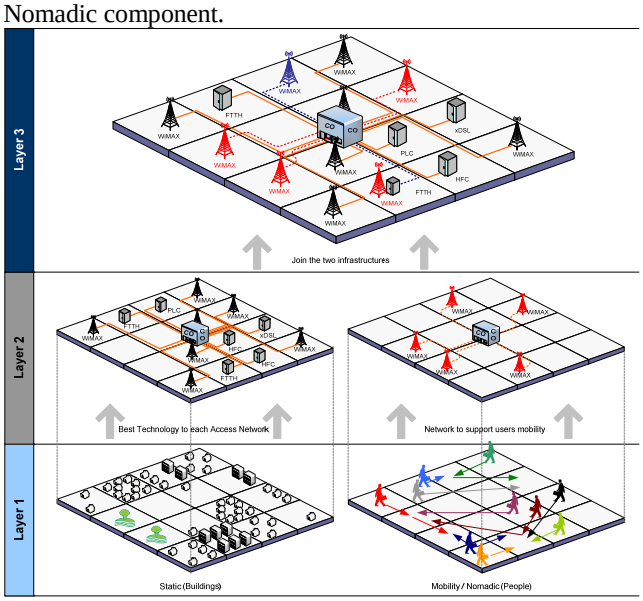


Figure 3. Cost model framework [11]

The Capital Expenses (CAPEX) costs referred above are divided into: equipment costs, installation costs, cable costs, housing costs and civil works. Besides the annual capital costs, which are derived from the relevant values for directly and indirectly attributable investments, other costs also need to be taken into account, for instance those incurred for the network's operation and maintenance (OPEX -Operational Expenses).

III. SEGMENTED REGULATION

A. *Overview*

One of main goals of regulated access is to prevent the incumbent from abusing a dominant market position. It is necessary make sure that alternative operators can compete effectively. It is fundamental that incumbent operators give access to the civil works infrastructure, including its ducts, and to give wholesale broadband access (bitstream) to the local loop (be it based on copper, new fiber, etc.). However, at the same time alternative operators should be able to compete on the basis of the wholesale broadband input while they progressively rolling out their own NGAN

 ISBN:978-1-61208-113-7

infrastructure. In some areas, especially with higher density, alternative operators have rolled their own infrastructure and broadband competition has developed. This would result in more innovation and better prices to consumers.

Many Europeans incumbents and some alternative operators are starting to plan and in some cases deploy large scale fiber investments, which results in important changes for European fixed line markets [3]. The risk of alternative operators will take longer to deploy their own infrastructure will give to incumbent the possibility to create new monopolies at the access level. The technologies used and the pace of development vary from country to country according to existing networks and local factors. Based on the different underlying cost conditions of entry and presence of alternative platforms, it may be more appropriate to geographically differentiate the access regulatory regime.

This part of the work focus the development of a tool using c language (with multiprocessing) that simulates the impact of retail and wholesale prices in provider's profit, welfare, consumer surplus, costs, Market served, network size, etc.

B. Description

In the proposed model "Retail Prices" represents the set of retail prices charged by providers for each service to consumers in a given region/area. We assume that retail providers cannot price discriminate in the retail market. "Wholesale Prices" represents the prices that one provider charges to other provider to allow the later to use the infrastructure to reach consumers. We assume that wholesale price can be different in each area. Also, we assume that when a provider buy infrastructure access in the wholesale market, cannot resell to another provider. The shared infrastructure consists of (Table I): Conduit and collocation facilities; Dark fiber leasing (dark fiber requires active equipment to illuminate the fiber – for example repeaters); and Bit stream.

TABLE I. INFRASTRUCTURE LAYERS (MAIN COMPONENTS)

Infrastructure Layer	Access Network: CO	Feeder Network	Street Cabinet	Secondary Network	Customer
Conduit and collocation facilities		Trenching and Ducting	Cabinet / Closure	Trenching and Ducting	
Dark Fiber		Fiber cable		FTTH: Fiber xDSL: Copper HFC: Coaxial PLC: LV Current WiMAX: ---	
Equipment	OLT Ports and Chassis Primary Splitter	Optical Repeaters	FTTH: Secondary Splitter xDSL: ONU and DSLAM (*) HFC: ONU and RF Combiner PLC: ONU and Coupling Unit (**) WiMAX: ONU and BS (***)	FTTH: Optical Repeaters xDSL: Copper Regenerator HFC: RF Amplifier PLC: Repeater for LV Network WiMAX: ---	FTTH: ONU xDSL: DSL Modem and Splitter HFC: Cable Modem PLC: PLC Modem WiMAX: WiMAX Modem

(*) DSLAM include: Line-Cards, Splitter, Chassis, and Racks
(**) Local MV/LV transformer station equipment include: O/E converter device at a pole or ground, and coupling unit (injection point)
(***) Base Station include: Antenna masts, PMP equipment (multiplexer + total sectors)

For example, Wholesaler provider can sell Layer 0 access (conduit and collocation facilities) and/or Layer 1 access (dark fiber leasing) or Layer 2 access (bitstream – network layer unbundling – UNE loop) only to retail providers and not directly to consumers. UNE loop is defined as the local loop network element that is a transmission facility between the central office and the point of demarcation at an end-user's premises.

Providers incur in fixed costs to build network infrastructure to provide access to a region and in marginal costs to connect each consumer separately. As we can see in Figure 4, our tool has several input parameters (one of them come from the techno-economic model described in the previous section), compute several results and find the strategies that are Nash equilibrium. The results are represented in tables and graphics.

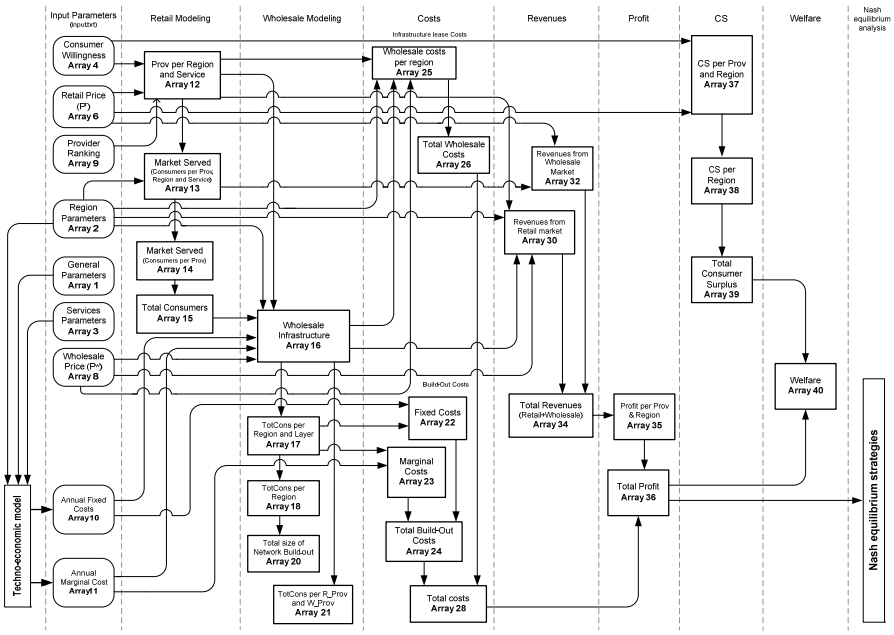


Figure 4. Simulation Tool architecture

The input parameters can be divided into 7 main groups:

- Identify the number of providers, regions, services and layers. We used layers to represents the shared infrastructure (Layer 0 – conduit and collocation facilities; Layer 1 – Dark Fiber leasing; and Layer 2 – Bitstream).
- Region parameters: for each region/area we need to define the total homes, Avg. Feeder length (central office (CO) – aggregate node (AGN)), Distribution Length (AGN - End User), and Geographical Area Description (Rural, Sub-urban, or Urban).
- Service parameters: Bandwidth required for each Service in the different regions.
- Willingness to pay for services per Region: Consumers have different willingness to pay for each service (voice, video and data). So, is necessary the definition of the willingness for each region and each service.
- Providers Retail Prices for the different services. One assumption is that consumers always buy one service when their willingness to pay is higher than the price at which some provider sells service. If there are two or more providers, consumers choose the service from the

provider with the lowest price. If several providers have the same price, then we use the provider ranking.

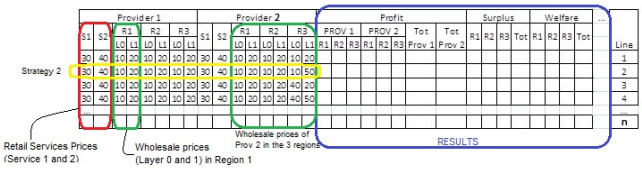
- Providers Wholesale Prices for the different layers in the several regions. We assume that each part of the infrastructure can had different leased prices in the each region.
- Fixed and Marginal costs. These costs are computed in the tool presented above. The cost model uses the parameters defined previous to compute the costs. For example: total homes, length CO-AGN, length AGN-End User, bandwidth required for each Service, etc.

Based on the several input parameters described, our tool computes several results (profit, consumer surplus, welfare, market served, network size, costs, and revenues) and finds the strategies that are Nash equilibriums.

1) Results

Next table show the structure of the results correspondent to a scenario of 2 providers, 2 retail services, 2 infrastructure layers (Layer 0: Conduit; Layer 1: Cable + Equipment) and 3 regions. Each line is a strategy (We consider a strategy a set of retail and wholesale prices)

TABLE II. STRATEGIES AND RESULTS (TOOL OUTPUT)



For each combination of prices, the tool computes: Profit, Consumer Surplus, Welfare, Market Served, Network Size, and Total Costs. The results are presented in several graphics (see Figure 5 and Figure 6).

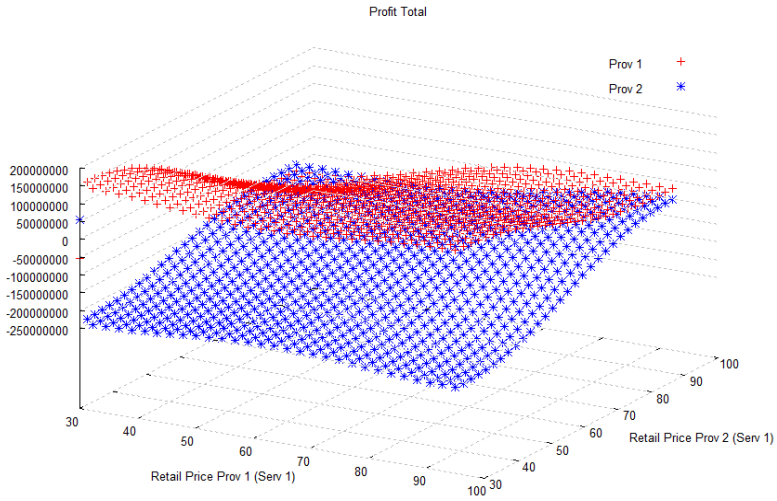


Figure 5. Total Profit (Retail Price)

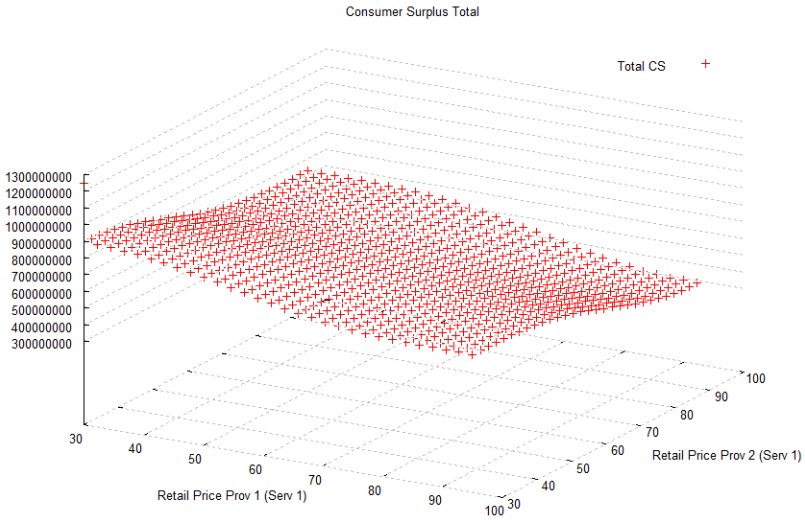


Figure 6. Total Consumer Surplus

2) Nash equilibriums

Nash equilibrium is a fundamental concept in the theory of games and the most widely used method of predicting the outcome of a strategic interaction in the social sciences. A game (in strategic or normal form) consists of the following three elements: a set of players, a set of actions (or pure-strategies) available to each player, and a payoff (or utility) function for each player. The payoff functions represent each player's preferences over action profiles, where an action profile is simply a list of actions, one for each player. A pure-strategy Nash equilibrium is an action profile with the property that no single player can obtain a higher payoff by deviating unilaterally from this profile (International Encyclopedia of the Social Sciences – 2nd Edition).

The next paragraphs explain the algorithm used in our work for finding Nash equilibriums - Finding each provider's best response to the other provider's strategy: 1) Select a specific provider and a specific strategy. For example, select strategy A from provider 1. Next, find the provider 2 best response (column 4) for the strategy A from provider 1. This means that the best response of provider 2 to strategy A from provider 1 is 6 (see Figure 7 – step 1). 2) For provider 2 select strategy A (the same strategy selected in step 1), and find the provider 1 best response (column 3) for the strategy A from provider 2 (see Figure 7 – step 1). 3) Repeat step 1 and 2 for strategy B. 4) When we finished, we will get the following table. Any line with a box in column 3 and 4 is a Nash equilibrium. In other words, when both providers are playing their best response at the same time, that is a Nash equilibrium (provider 1 plays strategy B and provider 2 plays strategy B) (see Figure 7 – step 3).

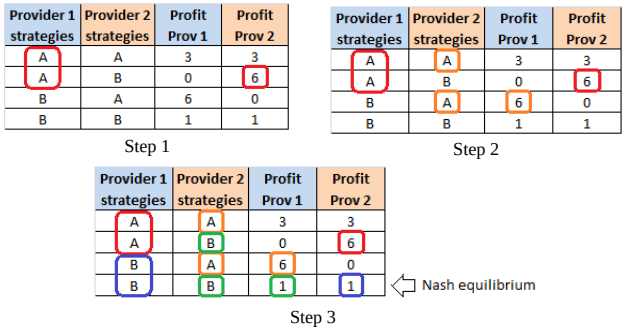


Figure 7. Finding Nash equilibrium (steps)

For the scenario presented above, our tool identified the combinations that are Nash equilibrium.

TABLE III. NASH EQUILIBRIA

Strategies from Provider 1								Strategies from Provider 2									
Serv 1	Serv 2	Region 1		Region 2		Region 3		Serv 1	Serv 2	Region 1		Region 2		Region 3		Line	Strategy
		Layer 0	Layer 1	Layer 0	Layer 1	Layer 0	Layer 1			Layer 0	Layer 1	Layer 0	Layer 1	Layer 0	Layer 1		
95	125	10	20	40	50	10	20	95	125	10	20	10	20	10	20	52417	1
95	125	10	20	40	50	10	20	95	125	10	20	10	20	10	50	52418	2
95	125	10	20	40	50	10	20	95	125	10	20	10	20	40	20	52419	3
95	125	10	20	40	50	10	20	95	125	10	20	10	20	40	50	52420	4
95	125	10	20	40	50	10	20	95	125	10	20	10	50	10	20	52421	5
95	125	10	20	40	50	10	20	95	125	10	20	10	50	10	50	52422	6
95	125	10	20	40	50	10	20	95	125	10	20	10	50	40	20	52423	7
95	125	10	20	40	50	10	20	95	125	10	20	10	50	40	50	52424	8
95	125	10	20	40	50	10	20	95	125	10	20	40	20	10	20	52425	9
95	125	10	20	40	50	10	20	95	125	10	20	40	20	10	50	52426	10
95	125	10	20	40	50	10	20	95	125	10	20	40	20	40	20	52427	11
95	125	10	20	40	50	10	20	95	125	10	20	40	20	40	50	52428	12
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	....	...
95	125	40	50	40	50	40	50	95	125	40	50	40	50	10	50		n

IV. CONCLUSIONS

The definition of the access price is a critical question when the incumbent also has activity in the retail market. Without the access price regulation the operator owner of the network can obstruct or hinder the access to the network. But, by other side, the exaggerated control of the access price can discourage the investments of the incumbent in the network quality. The investment in the network quality increases the services value to the existent consumers (for example: access with higher quality and speed to services like e-mail, www, video, etc.) and attracts new consumers.

So, regulatory authority can use the access prices definition to: induce the entrance of new providers in the retail market- concurrency in the retail market; incentive to investment; and consumer's welfare. The regulatory authorities need to define the max price for wholesale access. So, the main decisions are: Decision about the wholesale access price for each layer; and Decision about the price for each service in retail market.

The tool pretends support regulatory authorities to determine whether it is appropriate to delineate markets more narrowly than on a 'national' basis, and if so, how they should be segmented is a complex task. The experience of several OECD countries demonstrates that identifying relevant criteria for the definition of geographic markets and the segmentation of the market is possible and can be effective, but can be complex [2]. It is important that regulators accurately determine whether geographically segmented regulation is appropriate. When effectively implemented, geographic segmentation will promote competition and investment and serve the long-term interests of end-users. In addition, it may make sense for some countries to utilize geographically segmented regulation and for others to decline to do so. Also, to the extent that different technologies have different geographic footprints, the possibility arises that this could lead to distortions if different technologies of increasing substitutability (because of convergence) are regulated differently under a geographic regulation regime.

It is possible that investment in areas which remain regulated (e.g., sparsely populated rural areas) will be adversely affected by geographic regulation. This is because the incumbent's priority could become investment in areas open to competition to enhance its competitive prowess, and this could, in turn, result in competitive operators also focusing more attention to these areas rather than in rural areas.

REFERENCES

[1] F. Kirsch, C.V. Hirschhausen, Regulation of Next Generation Networks: Structural Separation, Access Regulation, or no Regulation at all?, in: First International Conference on Infrastructure Systems and Services: Building Networks for a Brighter Future (INFRA), Rotterdam, The Netherlands 2008, pp. 1-8.

[2] P. Xavier, Geographically Segmented Regulation for Telecommunications, in: Working Party on Communication Infrastructures and Services Policy, OECD, 2010, pp. 77.

[3] G.B. Amendola, L.M. Pupillo, The Economics of Next Generation Access Networks and Regulatory Governance in Europe: One Size Does not Fit All, in: 18th ITS Regional Conference, Istanbul, Turkey, 2007.

[4] R.D. Vega, NGANs and the geographical segmentation of markets, in: OECD Workshop on Fibre Investment and Policy Challenges, OECD, Stavanger, Norway, 2008, pp. 1-18.

[5] P. Ferreira, Modeling Segmented Regulation for Next Generation Networks, in: The 36th Research Conference on Communication, Information and Internet Policy, George Mason University School of Law, Arlington, VA, USA, 2008, pp. 1-29.

[6] J.P. Pereira, A Cost Model for Broadband Access Networks: FTTx versus WiMAX, in: 2007 Second International Conference on Access Networks (AccessNets '07), Ottawa, Ontario, Canada, 2007, pp. 1-8.

[7] E. Richards, Future broadband - Policy approach to next generation access, in, Ofcom, 2007, pp. 1-34.

[8] F. Castelli, C. Leporelli, Segmented regulation in global oligopolies: industry configuration and welfare effects, Inf. Econ. Policy, 7 (1995) 303-330.

[9] U. Stumpf, Towards geographical differentiation of broadband regulation?, in: 3rd Black Sea and Caspian Regulatory Conference, Istanbul, 2008, pp. 1-16.

[10] J.P. Pereira, The Role of WiMAX Technology on Broadband Access Networks, in: U.D. Dalal, Y.P. Kosta (Eds.) WIMAX, New Developments, IN-TECH, Vienna, Austria, 2009, pp. 17-45.

[11] J.P. Pereira, P. Ferreira, Access Networks for Mobility: A Techno-Economic Model for Broadband Access Technologies, in: TRIDENTCOM 2009 - The 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, IEEE, 2009.

 ISBN:978-1-61208-113-7

Does Cloud Computing Matter?
Networking IT and Services Value in Organizations

Cheng-Chieh Huang
Dept. of Information Management
National Taiwan University,
Taipei, Taiwan
d94725007@ntu.edu.tw

Ching-Cha Hsieh
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
cchsieh@im.ntu.edu.tw

***Abstract*—This article argues that cloud computing matters through interactions between organizations, IT, and cloud services. It illustrates the cloud computing value generation processes of Amazon, Google, IBM, and Microsoft and examines their strategies. Furthermore, this paper proposes value networking concepts and an ANT lens for future research on cloud computing and business values.**

Keywords-cloud computing; IT value, actor network theory

I. INTRODUCTION

Since Carr [1] published his article "IT Doesn't Matter" in the Harvard Business Review, scholars have started to reconsider the business value of information technology. Carr [1] argued that IT has become a kind of commodity like water or electricity, and is thus no longer specific strategic value to enterprise. The introduction of cloud computing partly reinforces Carr's claims, by potentially outsourcing certain IT functions to a third party service. However, cloud computing is also considered as a strategic weapon, helping enterprises lower the costs and increase their competitiveness. Does cloud computing matter or not?

Past literature on IT and business value divides IT value into two types. One, called technology determination, views IT as a strategic resource or innovative tool, and states that a specific IT can create value for organizations. Another, organization determination, claims that IT increases competitiveness when it's aligned with organizational strategy.

With this perspective, cloud computing appears to offer technical innovation, but not all organizations can enjoy the benefit immediately. It seems that IT and organizational value generation do not share a simple causal relationship [2].

Furthermore, cloud computing is not only a technological innovation but also a service innovation. Thus, any evaluation that fails to consider the service advantages of cloud computing neglects an important characteristic.

In this article, we consider the IT business value generated through dynamic interactions of organizations, IT artifacts and services. We use the cloud computing development cases of Amazon, Google, IBM, and Microsoft to demonstrate how different business values emerged through the dynamic interactions within these companies.

We argue that while IT is more service-oriented, a network view is needed to fully understand the relationship between IT, services, and business values. As a result, we employ actor network theory (ANT) as a lens to illustrate the value of cloud computing and the implications of further research.

In the following sections, we first review the literature of cloud computing and IT business value, and then propose an analytical framework. Next, we describe our methodology and use our research framework to illustrate four business case studies. Finally, we discuss our conclusions and identify contributions, limitations and suggestions for future research.

II. LITERATURE REVIEW

A. Cloud Computing

Cloud Computing generally refers to applications or IT resources delivered as services over the Internet, and the datacenter hardware and system software that provides those services. Definitions of cloud computing are diverse [3, 4]. Vaquero et al. provide one careful definition [4]:

Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically re-configured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized SLAs (p.51).

From this definition, cloud computing functions not only as an enabling technology but also as a service model. From the beginning, cloud computing has co-evolved as both a service and technology innovation (see Figure 1).

This implies that evaluations of cloud computing such as Carr [1] cannot only examine its IT characteristics, but also its service of economic model. Evaluations of cloud computing must consider both aspects.

 ISBN:978-1-61208-113-7

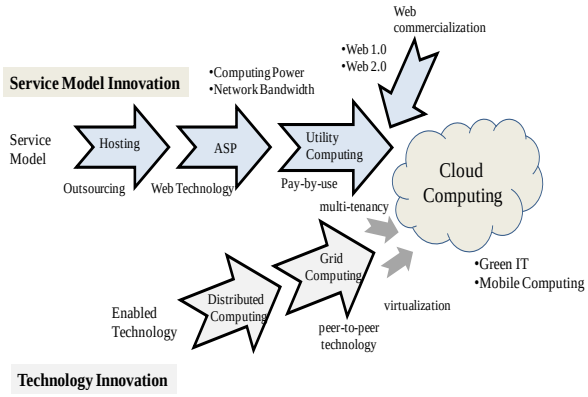


Figure 1. Co-evolution of Cloud Computing

B. *IT and Business Value*

Past literature on IT and business value focuses on three perspectives: assets, investments, and alignment (see Table 1). The asset perspective views IT as a strategic resource or an innovation tool that when introduced to organizations generates value. For example, Lynntien argues that Internet computing technology brings the disruptive nature of innovation to organizations [5]. Additional studies claim that IT generates business value through its combination with other complementary organizational resources, such as human resources or business relations [6, 7].

The investment perspective looks past organizational characteristics or strategy and instead focuses on financial models, such as the real option model or productivity factor counting [8, 9]. This perspective is problematic however, as existing literatures increasingly argues that organizational characteristics influence IT investment and firm performance relations [10].

The alignment perspective considers how IT may improve a firm's performance by fitting certain organizational needs or aligning with organization strategy [11][12]. However, this viewpoint largely fails to explain why certain Internet characteristics or cloud computing increase opportunities to strengthen a company's competitiveness.

TABLE I. LITERATURE REVIEW OF IT AND BUSINESS VALUE

Perspectives	IT and Business Value	Literature
assets	IT as strategic resource	[5] [6] [7]
	IT as innovation tool	
investments	IT as investment	[8] [9] [10]
	IT as productivity factor	
alignment	IT organizational fit information processing	[11] [12]
	IT aligned to business strategy	

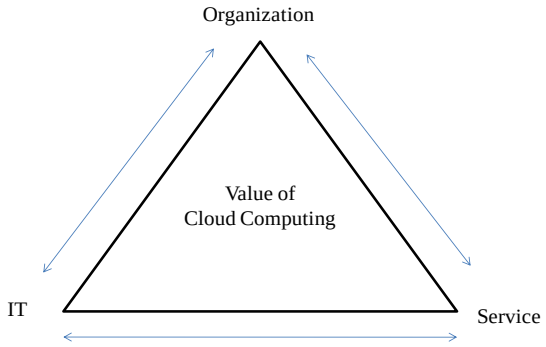


Figure 2. Research Framework

In summary, the potential relationship between IT and business value generation is not straightforward, but rather emerges through interactions between organization and IT [2].

C. *Research Framework*

Based on the literature described above, we design a framework for examining the cloud computing value generation process within different organizations. A diagram of this framework is presented in Figure 2.

III. METHODOLOGY

In this paper, we use case study methodology [13] to examine cloud computing and business value. We selected four firms: two Internet service firms (Amazon, Google), and two technology vendors (IBM, Microsoft). These four firms are famous for their use of cloud computing. Our data sources include documentation on their cloud computing development histories, news reports, company reports, successful cases, and independent analysis reports such as IDC, Gartner, and Ovum [14]. We also interviewed high-level managers to discuss their strategies and their perceptions on the values of cloud computing. All interviews were recorded. From this data, we use event analysis and our research framework to understand their value generation processes.

IV. CASE STUDIES

A. *Amazon*

Headquartered in Seattle, Amazon was established in 1994 as primarily an online bookstore. They soon expanded to flowers, software, electronic goods, toys, and eventually general retail items. Amazon was the first top 500 online retail business in the United States, with recorded profits in 2009 of roughly 24 billion dollars.

 ISBN:978-1-61208-113-7

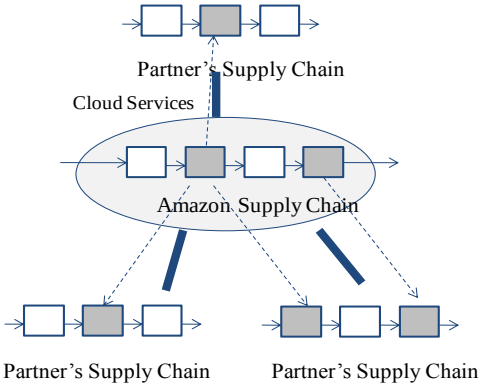


Figure 3. Amazon's Cloud Comuting Strategy

The development of cloud services at Amazon began in 2003, when it first offered web services for its e-commerce partners. For example, partners who used Amazon's online store to sell music CDs could use Amazon's services to rank the latest music purchases and examine customer comments in order to better market and sell their products. These web services helped partners use Amazon as a promotional site for their goods.

To better facilitate their small electronic store partners, Amazon gradually transferred their internal IT infrastructure to cloud services. This included such functions as storage (S3), server computing resources (EC2) and even e-commerce business processes, such as fulfillment processes (FWS), payment processes (FPS), and personnel matching processes (Mechanical Turk).

Through the development history of cloud computing at Amazon, we understand that Amazon first offered website design and development tools to help partners sell goods through Amazon's online store. After integrating various kinds of services with partners, Amazon strengthens the competitiveness of its whole supply chain operation (see Figure 3).

B. *Google*

The largest online Internet search engine in the world, Google is headquartered in California and was established in 1998. Relying on advertising revenue from its search engine business, Google earned 23.6 billion dollars in 2009. For several years now, Google has moved beyond online searching, as represented by its acquisition of YouTube, the development of the Android open source operating system, the Google Chrome Internet browser, Google Earth, and various cloud services.

Google first announced its cloud services in 2005. The primary purpose of the Google API is to let consumers log into their websites frequently, increasing web traffic, and thus encouraging advertisers to place their ads on Google websites.

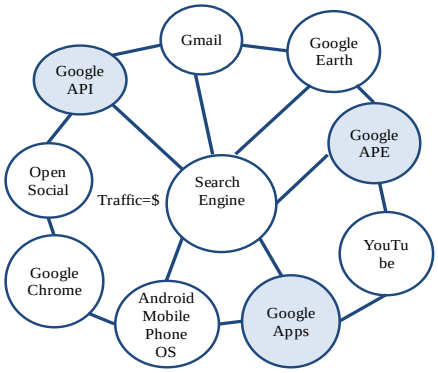


Figure 4. Google's Cloud Comuting Strategy

Later, Google developed various types of cloud services for both consumers and website designers, such as Google Docs, Google Financial, Google Spreadsheets, Google APE, and so fourth.

For Google, search engine traffic is revenue (traffic=$). That is, services are developed and branches are merged to help improve traffic. For example, YouTube or Open Social API were acquired and merged or linked with other social community websites in order to increase their popularity and thus increase traffic. Through the new Android operating system and Google Chrome, Google hopes to connect cell phones and browsers directly to its search service, and make this service more convenient.

As Figure 4 shows, cloud services support Google's "traffic equals money" strategy, which attracts more consumers to its search engine and thus increases advertising revenue.

C. *IBM*

Established in 1924, IBM started with making enterprise information hardware, such as electronic calculators, large-scale mainframes, and the first generation of personal computers. Recently, IBM has shifted its business towards services and software provided to large enterprises.

IBM started developing cloud services to help its small independent software vendor (ISVs) partners located worldwide use IBM's servers or storage capacity. This obviated the need for ISVs to invest in hardware/software, and allowed them to develop software through IBM's own platform. Later, IBM developed their cloud computing technology into products that support their large enterprise customers in building their own cloud data center. IBM's online cloud services help showcase their cloud computing technology solutions.

 ISBN:978-1-61208-113-7

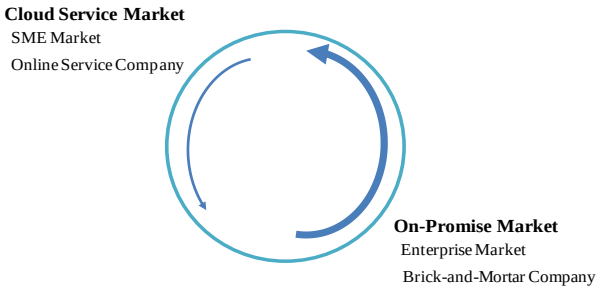


Figure 5. IBM's Cloud Comuting Strategy

IBM attempts to use the cloud computing technology products and leverage their consulting services and software implementation experiences in the large-scale enterprise and then explores to small and medium enterprises and on-line service companies market.

Take her cloud services implementation experiences in UPS for example, IBM combined cloud services with their software implemented in UPS. IBM supported their customers, UPS and also touched UPS's online partners. It is so-called two-sided market strategy includes the large-scale enterprise software service market (on-promise market) that IBM has already deeply engaged and new developing medium and small-scale online service companies (cloud service market).

For IBM, cloud services and technology play a bridge role to explore on-line and small medium enterprise (SME) markets (see Figure 5).

D. *Microsoft*

Established in 1983, Microsoft was an early leader in computer operating systems and suite software on the personal computer with both its MS-DOS operating system and MS-Office software suite.

Microsoft earned 58 billion dollars in profit in 2009. Its personal and commercial Office series accounts for more than 90% of the market.

Despite its dominant market position, Microsoft realizes the growing trend towards online services, and that PC or on-promise software are no longer the only choices. It is thus finding ways to combine its software expertise with online services.

This is the concept of "software plus services" or "3 screens and a cloud" that Microsoft announced in 2009. For Microsoft, cloud services or cloud computing technology helps the company smoothly transition to a new "network operating system" by combining their traditional on-promise software with these services (see Figure 6).

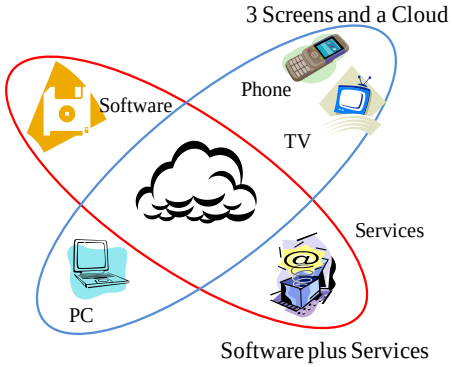


Figure 6. Microsoft's Cloud Computing Strategy

V. LANALYSIS AND DISCUSSION

A. *Networking IT/Service Value*

The four cases described above demonstrate that different companies view different opportunities with cloud computing, and align their strategies accordingly to generate value.

For these firms, cloud computing is not only a technology artifact, but also a part of their service model. In this way, it represents a techno-economic network (TEN) [15].

Callon described TEN as "a coordinated set of heterogonous actors which interact more or less successfully to develop, produce, distribute and diffuse methods for generating goods and services." From Callon's point of view, the economic value is generated from actors, intermediaries (nonhuman), translation and their relationships [15].

In Amazon's case, cloud computing services stemmed from their internal IT and originally supported their business processes. Then, Amazon enrolled their e-commerce partners, adopting their web services, and then embedding their cloud services within their daily business operations. Amazon thus used IT and cloud services to form their partners' networks and strengthen their own business value (Figure 7).

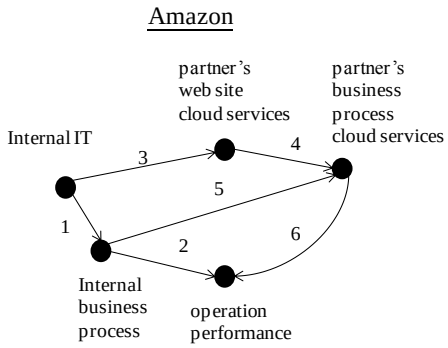


Figure 7. Amazon's Value Networking

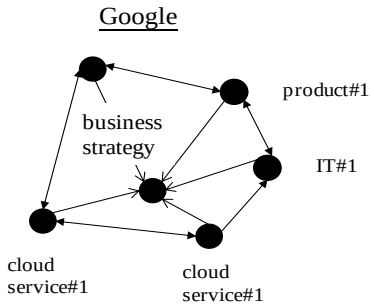


Figure 8. Google's Value Networking

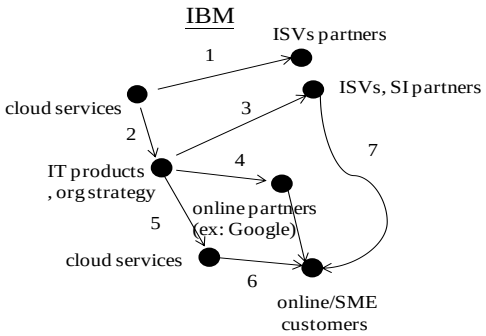


Figure 9. IBM's Value Networking

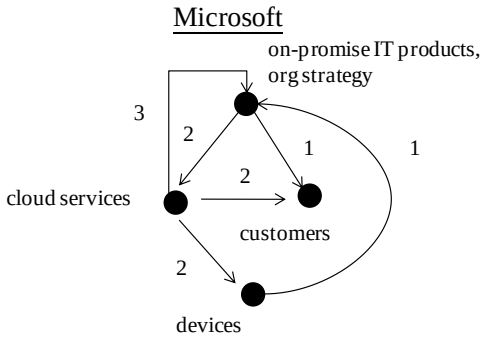


Figure 10. Microsoft's Value Networking

In Google's case, their cloud services, IT products, and tools all support their advertising revenue business strategy. Moreover, these services and products leverage each other and thus intensify the value of their whole network (see Figure 8).

In IBM's case, cloud services originally helped IBM's small independent software vendor (ISV) partners leverage IBM's software/hardware resources through the Internet. Then, IBM strengthened and transferred their technology products to the cloud in support of their secondary market: online/SME customers. Clouds services acted as a bridge to a new market network [6] (see Figure 9).

In Microsoft's case, Microsoft leveraged cloud services to complement their on-promise software, and strength their product's value. Microsoft then hopes to enroll its customers into this new value network (see Figure 10).

According to our analysis, IT/services generate business value dynamically, and thus not only through strategy alignment or IT determination, as suggested by previous literature. Value is generated through networking activities between heterogeneous actors and IT/services. We refer to this as "Networking IT/service Value."

With cloud computing, internal IT artifacts are transferable to cloud services (as with Google) or strengthened by cloud services (as with Microsoft). However, in considering IT value, we should both discuss the properties of IT artifacts and the service models they serve.

B. ANT Lens and Cloud Computing Value

Actor Network Theory (ANT) was developed in the sociology of science and technology [16]. ANT helps describe how actors form alliances, involve other actors and use non-human actors (artifacts) to strengthen such alliances and secure their interests. ANT consists of two concepts: inscription and translation.

Inscription describes what characteristics an engineer designs, develops, and diffuses into a technical artifact. For translation, when an actor-network is created, it consists of four translation processes [17]:

- problematization: The focal actor defines the interests that others may share, establishes itself as indispensable, and sets the obligatory passage point through which all the actors in an actor-network must pass.
- interessement: The focal actor convinces other actors.
- enrollment: Other actors accept the interests as defined by the focal actor.
- mobilization: The focal actor uses a set of methods to ensure that the other actors act according to their agreement and do not betray this agreement.

Regarding the cases in this study, what do these companies inscribe into their cloud computing technology or services? For example, Amazon designed its cloud computing technology and services for strengthening its e-commerce partners' business processes. Are the properties of a technology or service model transferable to a brick-and-mortar enterprise?

Second, how do our case companies mobilize their partners to join the network? How do they set the obligatory passage point through which all actors pass?

In Table 2, we analyzed the translation and inscription characteristics of our four case companies' cloud computing actor-networks.

ANT assumes the properties of actors or non-humans are static. However, in the case of cloud computing, IT artifacts may transfer to services, and services also can strengthen IT artifacts. These dynamic interactions and transformations need to be considered in future research.

 ISBN:978-1-61208-113-7

TABLE II. TRANSLATION AND INSCRIPTION OF CLOUD COMPUTING

Focal Company	Translation	Inscription
Amazon	Problemaitziaion, Interessement, Envrollment Mobilization	efficient supply chain process
Google	Problemaitziaion, Interessement, Envrollment	online services
IBM	Problemaitziaion, Interessement	smart services connection
Microsoft	Problemaitziaion, Interessement	software plus services

TABLE III. CLOUD COMPUTING VALUE RESEARCH IMPLICATIONS THROUGH ANT LENS

ANT Lens	Research Implications
Inscription	1. Does the originate design of cloud computing technology/service impact their networking and value generation?
Translation	1. What are actors' translations of cloud computing? 2. How do they negotiate their interests of cloud computing?
Obligatory passage point	1. What are the obligatory passage points of different actors? 2. Are the obligatory passage points different in every actor-network?
Technology/service	1. How the non-humans (IT or services) convert each other characteristics impact actor-network?
Competitive networks	1. Why and how do these companies generate different network values? 2. How do these different actor-networks compete or collaborate?

Finally, in our case analysis, these companies generate different value networks. Do these networks compete? How do they compete?

In Table 3, we list additional issues and implications that are interesting for further research on the value of cloud computing, as seen through the ANT lens.

VI. CONCLUSION

In this article, we discuss cloud computing value generation through four company case studies. We argue that cloud computing value is generated through the interactions of organization, IT and services. We further analyze their value networking activities, and then propose ANT as a lens to interpret the cloud computing value generation process. ANT provides research implications for further research on cloud computing value and competitive strategy.

The analysis of this paper is limited to the four major company cases. Future research may investigate smaller firms, and analyze firm activity, actor responses, and their translation within the actor network in more detail.

REFERENCES

[1] N. G. Carr, "IT Doesn't Matter", Harvard Business Review, 2003, pp. 41-49.

[2] M. L. Markus and D. Robey, "Information Technology and Organization Change: Casual Structure in Theory and Research", Management Science, 1988, pp. 583-598.

[3] M. Armbrust and UC Berkeley RADSL, "Above the Clouds: A Berkeley View of Cloud Computing", http://radlab.cs.berkeley/edu/, 2009, pp. 1-23.

[4] L. M. Vaguero, and L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Toward a Cloud Definition", ACM SIGCOMM Computer Communication Review, 2009, pp. 50-55.

[5] K. Lyytinen and G. M. Rose, "The Disruptive Nature of Information Technology Innovations: The Case of Internet Computing in Systems Development Organizations", MIS Quarterly, 2003, pp. 557-595.

[6] N. Melville, K. Kraemer, and V. Gurbaxani, "Review: Information Technology and Organization Performance: an Integrative Model of IT Business Value", MIS Quarterly, 2004, pp. 283-322.

[7] T. Ravichandran and C. Lertwongsatien, "Effect of Information Systems Resources and Capabilities on Firm Performance: A Resource-Based Perspective", JMIS, 2005, pp. 237-276.

[8] M. Benaroch, S. Shah, and M. Jeffery, "On the Valuation of Multistage Information Technology Investments Embedding Nested Real Options", JMIS, 2006, pp. 239-261.

[9] P. E. D. Love and Z. Irani, "An Exploratory Study of Information Technology Evaluation and Benefits Management Practices of SMEs in the Construction Industry", 2004, pp. 227-242.

[10] L. Motiwalla, M. R. Khan, and S. Xu, "An Intra- and Inter-Industry Analysis of E-business Effectiveness", I&M, 2005, pp. 651-667.

[11] J. F. Fairbank, G. Labianica, H. K. Steensma, and R. Metters, "Information Processing Design, Choices, Strategy, and Risk Management Performance", JMIS, 2006, pp. 293-319.

[12] T. A. Byrd, B. R. Lewis, and R. W. Bryan, "The Leveraging Influence of Strategic Alignment on IT Investment: An Empirical Examination", Information and Management, 2006, pp. 308-321.

[13] R. K. Yin, Case Study Research, Design and Methods, 2nd Edition, Sage Publications, CA, 1994.

[14] G. McCulloch, Documentary Research in Education, History and the Social Sciences, Routledge Falmer, London, 2004.

[15] M. Callon, "Techno-economic Networks and Irreversibility", in: Law J. (ed.), A Sociology of Monsters: Essays on Power, Technology, and Domination, Routledge, New York, 1991, pp. 132-164.

[16] M. Callon, and B. Latour, "Unscrewing the big Leviathan", in: Knorr-Cetina, K., Cicourel, A.V. (eds.), Advances in Social Theory and Methodology. Routledge & Kegan, London, 1981, pp. 277–303.

[17] M. Callon, "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St. Brieuc Bay", in: Law, J. (ed.), Power, Action and Belief, Routledge and Kegan Paul, London, 1986, pp.197–233.

 ISBN:978-1-61208-113-7

Motivations and Challenges of Global Mobility with Universal Identity: A Review

Walaa F. Elsadek, Mikhail N. Mikhail
Department of Computer Science and Engineering, the American University in Cairo,
P.O. Box 74, New Cairo 11835, Egypt
walaa.farouk@aucegypt.edu, mikhail@aucegypt.edu

***Abstract* — Researchers are directing enormous efforts to achieve the aim of global mobility by enhancing the standard mobile IP with various routing schemes focusing on best routes and least cost while ignoring the facts of the organizations usual use of private IP's and the presence of firewalls. Nevertheless, existing Mobile IP models are still missing three basic concepts that hinder their applicability in real environment. First, global mobility must be independent of the different infrastructure technologies (e.g., Wi-Fi, WiMAX, UMTS, etc.). Second, a secure authentication mechanism for guiding the access of mobile nodes to the corporate network's resources is certainly needed. Third, the capability of correlating the mobile node's activities to a real world identity is a requirement of security in a wider since i.e., network, web, and national security. This paper defines the role of global mobility in facilitating and improving mobile business performance. It, also, presents a review of literature for the existing standards and schemes of mobility and analyzes their limitations. Finally, a reference is made to a practical approach for secure global mobility without the current limitations.**

Keywords- Mobile Computing; Interworking; Mobility; Mobile IP; Security; Wireless.

I. INTRODUCTION

Mobility in the enterprise is derived by both technology availability and the increase in user demand. The merging of 4G and WLAN networks prompts the needs of operators to increase their coverage with a blended service offering that makes best use of their old investments in legacy infrastructure, low price technology and new technology at least price to address the higher volume of delivered rich data services [1]. Many people think of wireless and mobility as plumbing – focusing only on infrastructure and the fundamental technical security challenges, privacy, platform standardization, and legacy system integration [2]. However, the real target should be the ability to drive business improvement, and that requires vision in scoping mobility to fit the enterprise while preserving the privacy and security of mobile users accessing critical applications and identifying them with unique universal digital identities. For true global mobility, the following key features need to be emphasized:

- Seamless roaming between heterogeneous wireless, wired, and ad-hoc networks as illustrated in Figure 1.
- No restriction on the type of the hardware (mobile sets, PDA, laptop, etc.) or their operating systems.
- The connection between different mobile operators and internet service providers must be smooth without the need of complex reconfiguration.
- Enhanced security mechanism to facilitate the creation of e-commerce, banking services as well as any other services that need strong authentication.
- Transparency to end-user that does not need complex application or am increase in power consumption.
- Scalable routing mechanism that is flexible in adopting large-scale macro-mobility and local scale micro-mobility.
- Minimum handover interruptions to enhance availability and reliability of the services provided either to or by roaming clients.
- The capability of correlating the user activities to a unique universal digital identity.

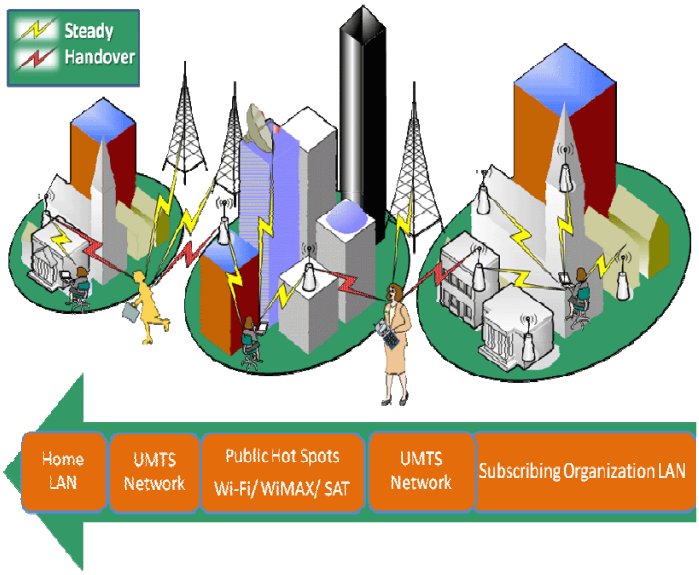


Figure 1. Overview of Global Mobility

II. MOTIVATIONS FOR GLOBAL MOBILITY WITH UNIVERSAL IDENTITY

A. Carriers' Motivations

Threats such as distributed-denial-of-service (DDoS) attacks, turbo worms, phishing, viruses and e-mail spam generates a huge amount of infected traffic that lead to subsequent outbreaks and disrupt the normal operation of a modern network. The primary challenges faced by today's service providers are maintaining service availability in the presence of such outbreak of malicious traffic. Security has become a critical characteristic of all services due to the direct effects reflected on the profit line of service providers. Universal digital identity can help in identifying the real identity of sources of threats then blocking their traffic or redirecting it to fake destinations.

Carriers are, also, challenged to meet the demand of their subscribers of enhanced mobility. They have to do so while avoiding new huge investment in remote areas by signing service level roaming agreements (SLRA) with other operators to make use of other infrastructure.

B. Subscribers' Motivations

New devices and business practices, such as PDAs and the next-generation of data-ready cellular phones and services, are the driving interest in the ability of a user to roam while maintaining two-way network connectivity:

- Roaming employees need to remain connected to their home corporate accessing local resources without any change to the corporate security policy and with least cost.
- Corporate employees need to keep accessing resources or services hosted by their roaming colleagues independent of their physical location using the same local private addresses.

In addition, the Mobile Nodes should remain transparently accessible to any corresponding node. Corresponding nodes are able to keep using the same addresses and with no need to any additional software.

C. Governments' Motivations

Criminals and terrorists are migrating to the digital world as it is tremendously lucrative and has less risk. They are willing to commit identity fraud and eager to sell it for profit. Negative impacts are induced on trusted transactions in online commerce, cyber investigations, authenticated individuals, or organizations, who want to gain access to services, systems, and facilities. Protection of assets and critical infrastructure such as telecommunications, public health, and the power grid, are necessary for the functioning of society [3]. In battle, war fighters must be able to identify people and determine if they are friend or foe, as well as if and how much of a risk they present. The challenge is to provide the war fighter with real time accurate information [4].The resultant escalation of cyber crimes and cyber attacks has resulted in the need for improved cyber investigations, security, and cyber defense. Grouping the network activities by universal digital identities enhances the cyber crime investigations with a tool that can simply reveal the real-world identities.

1) The Cyber Threats [4]

- Account take-over fraud in banking sectors, retailer, and healthcare provider.
- Access fraud on credit and financial information.
- Identity fraud in thin file situations and attack on an identity database.
- Cyber threat to enterprise attribute-based controls.
- Internal abuse of corporate assets and information.
- Wrest or hijack identity using Zombie networks.

2) The Cyber Challenge [4]

- Cyber security includes data protection, fraud detection, and preventions.
- Policy management that relies on security policies legislation to protect from identity theft.
- Breach detection by monitoring unauthorized system access or data acquisition using intrusions detection systems.
- Tracing and monitoring the usage of identities to detect unauthorized usage.
- Strong authentication methodology that correlates the identity user to the real identity owner.

III. REVIEW OF STANDARD MOBILE IP IN IPV4

Mobile IP (MIP) is an open standard, defined by the Internet Engineering Task Force (IETF) RFC 3344. Mobile IP enables users to keep the same IP address while traveling to a different network, ensuring that a roaming individual can continue communication [5]. Mobile IP does not drop the network prefix of the IP address of the node. Consequently, IP routing will succeed to route the packets to the node after movement to the new link [6]. RFC 3344 is considered the base for MIP. It defines the various entities involved in Mobile IP protocol and how they interact together to enable the registration of a roaming mobile node (MN) to the home network thus the home agent can forward the packet destined to MN to its care-of-address obtained from the foreign network.

RFC 4721: "Mobile IPv4 Challenge/Response Extensions" updates RFC 3344 by including a new authentication extension called the Mobile-Authentication, Authorization, and Accounting (AAA) Authentication extension. This new extension enables a mobile node to supply credentials for authorization, using commonly available AAA infrastructure elements. This authorization-enabling extension may co-exist in the same Registration Request with authentication extensions defined for Mobile IP Registration by RFC 3344 [6].

RFC 3344 assumes that tunneling is required for packet from the home agent to the mobile node's care-of address, but rarely in the reverse direction. It assumes that routing is independent of the source address and MNs can send their packet through the router in the foreign network. This assumption is not valid. This raises a need to establish a topologically correct reverse tunnel from the care-of address to the home agent [7]. RFC 2344: "Reverse Tunneling for

 ISBN:978-1-61208-113-7

Mobile IP" proposes backwards-compatible extensions to Mobile IP in order to support topologically correct reverse tunnels. When the mobile node joins a foreign network, it listens for agent advertisements and selects a foreign agent that supports reverse tunnels. It requests this service when it registers through the selected foreign agent. At this time, and depending on how the mobile node wishes to deliver packets to the foreign agent, it also requests either Direct or Encapsulating Delivery Style.

- **In the Direct Delivery Style:** the mobile node designates the foreign agent as its default router and proceeds to send packets directly to the foreign agent, that is, without encapsulation. The foreign agent intercepts them, and tunnels them to the home agent.
- **In the Encapsulating Delivery Style:** the mobile node encapsulates all its outgoing packets to the foreign agent. The foreign agent decapsulates and re-tunnels them to the home agent, using the foreign agent's care-of address as the entry-point of this new tunnel.

The MIP RFC3344 standard falls short of the promise in fulfilling the need of an important customer segment, corporate users (using VPN for remote access), who desire to add mobility support to have continuous access to Intranet resources while roaming outside the Intranet from one subnet to another, or between the VPN domain (i.e., trusted domain) and the Internet (i.e., un-trusted domain). Both firewall and VPN devices typically guard access to the Intranet. The Intranet can only be accessed by respecting the security policies in the firewall and the VPN device. In addition, any solutions to be proposed would need to minimize the impact on existing VPN and firewall deployments [8]. IP-in-IP tunneling does not generally contain enough information to permit unique translation from the common public address to the particular care-of address of a mobile node or foreign agent, which resides behind the NAT; in particular, there are no TCP/UDP port numbers available for a NAT to work with. For this reason, IP-in-IP tunnels cannot in general pass through a NAT, and Mobile IP will not work across a NAT [9].

RFC 3591: "Mobile IPv4 Network Address Translation (NAT) Traversal" enables mobile devices in collocated mode that use a private IP address (RFC 1918) [10] or foreign agents (FAs) that use a private IP address for the care-of address (CoA) are able to establish a tunnel and traverse a NAT-enabled router with mobile node (MN) data traffic from the home agent (HA) [9]. However, if the network does not allow communication between a UDP port chosen by a MN and the HA UDP port 434, the Mobile IP registration and the data tunneling will not work. Only the IP-to-UDP encapsulation method is supported.

The need is increasing for enabling mobile users to maintain their transport connections and constant reach ability while connecting back to their target "home" networks protected by Virtual Private Network (VPN) technology. This implies that Mobile IP and VPN technologies have to coexist and function together in order to provide mobility and security to the enterprise mobile users. RFC 4093: "Mobile IPv4 Traversal OF Virtual Private Network (VPN) Gateways" addressed the previous limitation by forcing any MN roaming outside the Intranet to establish an IPSec tunnel to its home VPN gateway first, in order to be able to register with its home agent. This is because the MN cannot reach its' HA (inside the private protected network) directly from the outside. This implies that the MIPv4 traffic from the MN to a node inside the Intranet is forced to run inside an IPSec tunnel. This in turn leads to distinct problems depending on whether the MN uses co-located or non-co-located modes to register with its HA

In co-located mode, successful registration is possible but the VPN tunnel has to be re-negotiated every time the MN changes its point of network attachment, as the MN's IP destination address changes on each IP subnet handoff, IPSec tunnel needs to be re-established. This could have visible performance implications on real-time applications and in resource-constrained wireless networks [11].

In foreign agent care-of address, MIPv4 registration becomes impossible. This is because the MIPv4 traffic between MN and VPN gateway is encrypted, and the FA (which is likely in a different administrative domain) cannot inspect the MIPv4 headers needed for relaying the MIPv4 packets. The use of a 'trusted FA' that is actually a combined VPN GW and FA can work fine in this case, as the tunnel end-points are at the FA and the VPN gateway as shown in Figure 2.

Limitation:

(i) However, due to security limitation, this scenario is not realistic in the general mobility case. It is not expected that the FA in access networks (e.g., wireless hot spots or CDMA 2000 networks) will have security associations with any given corporate network to apply 'trusted FA'.

(ii) This solution would leave the traffic between FA and MN unprotected. This is clearly undesirable as this link in particular may be a wireless link

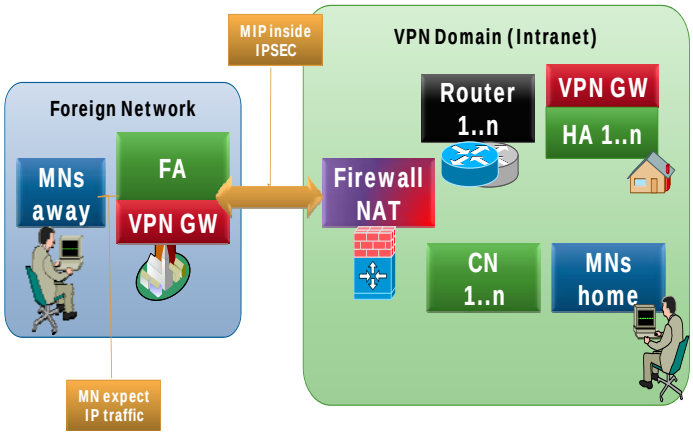


Figure 2. The use of a 'trusted FA' In foreign agent care-of address

IV. REVIEW OF STANDARD MOBILE IP IN IPv6

3775: “Mobility Support in IPv6” specifies a protocol, which allows nodes to remain reachable while moving around in the IPv6 Internet. This protocol allows a mobile node to move from one link to another without changing the mobile node's "home address". Packets may be routed to the mobile node using this address regardless of the mobile node's current point of attachment to the Internet. The movement of a mobile node away from its home link is thus transparent to transport and higher-layer protocols and applications [12]. A mechanism, known as "dynamic home agent address discovery" is added in Mobile IPv6 to provide support for multiple home agents and the home network reconfiguration. This mechanism allows a mobile node to dynamically discover the home agent IP addresses on its home link, even when being away from home. Mobile nodes can also learn new information about home subnet prefixes through the "mobile prefix discovery" mechanism.

A. Comparison between Mobile IPV4 and IPv6

1. There is no need to deploy special routers as "foreign agents", as in Mobile IPv4. Mobile IPv6 operates in any location without any special support required from the local router.
2. Support for route optimization is a fundamental part of the protocol, rather than a nonstandard set of extensions.
3. Mobile IPv6 route optimization can operate securely even without pre-arranged security associations
4. Support is also integrated into Mobile IPv6 for allowing route optimization to coexist efficiently with routers that perform “ingress filtering" [13].
5. The IPv6 Neighbor Unreachability Detection assures symmetric reachability between the mobile node and its default router in the current location.
6. Most packets sent to a mobile node while away from home in Mobile IPv6 are sent using an IPv6 routing header rather than IP encapsulation, reducing the amount of resulting overhead compared to Mobile IPv4.
7. Mobile IPv6 is decoupled from any particular link layer, as it uses IPv6 Neighbor Discovery instead of ARP. This also improves the robustness of the protocol.
8. The use of IPv6 encapsulation (and the routing header) removes the need in Mobile IPv6 to manage "tunnel soft state".
9. The dynamic home agent address discovery mechanism in Mobile IPv6 returns a single reply to the mobile node. The directed broadcast approach used in IPv4 returns separate replies from each home agent.

B. Summary of Mobile limitations

Unluckily, standard MIP faced many limitations that hindered its applicability in real environments as:

1. Ignoring that the Mobile Node (MN) can exist behind a Firewall or NAT device.
2. Considering only Forward Tunnel; that Corresponding Node (CN) packets has to be forwarded by Home Agent (HA) to MN while Ignoring Reverse Tunnel; that MN needs to access the corporate services.
3. With updated RFC 2344, that includes Reverse Tunnel; MIP becomes subjected to DoS and Session Hijacking due to the lack of a secure method for authenticating MN [7].
4. Ignoring the corporate security policy, during the tunnel establishment phase or the MN’s registration with the HA.
5. Ignoring that the DNS domain name is used for locating and addressing devices worldwide.
6. The VPN tunnel has to be re-negotiated every time the MN changes its point of attachment.
7. Ignoring the Scalability Factor; Home agent and Foreign agents are a single point of failure.

V. NEW MOBILITY ARCHITECTURE

In the paper: “Universal Mobility with Global Identity (UMGI) Architecture” [14], a new mobile IP concept has been introduced to address the limitations of previous proposed mobile IP. The UMGI correlates mobile users’ credentials as IP, hostname, and network equipment identifiers (ex. network cards or mobile sets) with their (U)SIM (Universal subscriber identity module) cards .Not only this corelation provides a strong method for authenticating subscribers but also it acts as the foundation of the newly proposed mobile IP protocol. The extracted “Country codes” and “Carrier Code” stored in the (U)SIM enable the dynamic discovery of the home agent thus facilitate routing the traffic to the home network. In addition, this paper has discussed the integration of the MN authentiation with the standard methods of authentionation in the UMTS network. This architecture extends the capabilities of standard mobile IP, solves its applicability problems, and links mobile users’ activities to unique universal identities

A. Advantages of the suggested UMGI:

1. Not like the standard MIP, which runs on the mobile node and keeps monitoring the network prefix, the UMGI Architecture makes the mobile node unaware of any procedure. All UMGI services and modules run on distributed servers administrated by the carriers or the corporate networks thus enhancing the performance of mobile subscriber while decreasing the battery consumption of the MN.
2. Standard MIP has not considered any handover procedure. It focuses on stationary hosts that moved to different location other than the home network. UMGI MIP inherits its mobility from the wireless technology adopted. It opens rooms for programmers to enhance the handover procedures thus movement can be unnoticeable.

 ISBN:978-1-61208-113-7

3. The security policy of the corporate network is preserved:
   a. By restricting the Home UMGI IPSec tunnel establishment to authorized the carrier gateways and the predefined UMGI Tunnel Subnet.
   b. The corporate firewall can be configured with an advanced security policy controlling the UMGI roaming subscribers' access and privileges.
4. DHCP Classification: preserves the UMGI roaming subscriber's CoA during the movement inside Hot Wi-Fi spots covering one or more buildings that have multiple APs but with a single border gateway and during the movement between multiple Node B or BTSs connected to the same access point on the GGSN. The UMTS or (E) GPRS access point can cover a country region.
5. Automatic private IP addressing procedure: Preserves the mapping of UMGI subscriber public address and private addresses as well as the DNS domain name while roaming in the foreign network. This enables MNs, in either foreign carrier, to handoff between multiple Wi-Fi hotspots and multiple UMTS or (E) GPRS access point even if the CoA obtained from the DHCP server is changed. The only challenge is the time required to update the mapping on the FG of the network to which the subscriber is attached. The FA should be capable to update the FG in few seconds. In TCP traffic, no packet will be lost as this will be regarded as congestion and retransmission will occur. For UDP, any packet lost in range of few seconds will not be noticed.
6. The combination of "UMGI Trust Relation" and the "Hierarchical Discovery Routing Procedure" increases the scalability and the security of the new mobile IP architecture while preserving the security of communication between foreign and home networks. The combination makes the architecture extremely customizable fitting small ISPs and large carriers. In addition, it increases the architecture flexibility to adopt several designs and different types of agreement starting from small ISPs inside the country and ending with regulators agreements cross-countries boundaries.
7. The carriers can freely add or modify the configuration of its gateway even the IP addresses without any need to update any other carrier under UMGI SLRA.
8. The architecture solves the current MIP applicability limitation. With UMGI, it is simple to create dynamic IPSec tunnel on demand with the home VPN gateway and to path through any firewall security policy.
9. Added a security layer that is boasted by a strong client authentication mechanism as EAP. This provides a strong protection against session hijacking and Denial of Service attack.
10. Using a single path between remote and local carrier, UMGI Remote Tunnel, to carry the MNs' traffic from multiple corporate networks connected to the same carrier gateway, decreases the UMGI subscriber's joining time by avoiding the process of "Tunnel setup Procedures" for MNs belonging to the same carrier.
11. The synchronization between both foreign and home carrier through AAA and HLR enforces the status consistency and increases the security by restricting access to only one UMGI subscriber per IMSI at time, even if having multiple registered equipments.
12. Mobile IP leaves transport and higher protocols unaffected. Other than mobile nodes/routers, the remaining routers and hosts will still use current IP address format without any modification. Unlike, Standard MIP, UMGI MIP does not need enabling jumbo frames or any change in the IP frame format. Thus, UMGI suits LAN/WAN topology.
13. UMGI is Multi-Vendor Interoperable. It can be considered as an organized setup of the standard protocols thus, no need to any software upgrade or any major change to the existing infrastructure.

VI. CONCLUSION

Enhancements to the standard Mobile IP techniques are being developed to improve mobile communications and to overcome the existing limitations by making the process more secure and more efficient. Researchers are continuously adding achievement to augment its applicability to the new business needs. In this paper, the importance of secure global mobility as motivated by the needs of corporate organizations, service providers, and governments is highlighted. State of the art schemes and standards dealing with facilitating and managing mobile computing both in the current IPv4 and the coming IPv6 are reviewed.

A reference is made to a suggested architecture that is aimed at overcoming the current practical limitations. As it should be, the suggested scheme is transparent, secure, scalable, independent of any communication protocol, and valid for hybrid infrastructure. This shows that designing an architecture for global mobility with universal identity that satisfies the requirements of service providers, governments, and corporate organizations is a challenge and needs an enterprise secure cooperations between the various entities. This architecture has proposed a new mobile IP that solves the standard mobile IP scalabilty limitations by proposing a solution that can be easily deployed with the presence of firewalls and VPNs. Also, the proposed solution has shown that a mobile user can handover hybrid infrastructure with very short delay without changing its real IP address or DNS domain name. Solving the challenges that hindered the applicability of standard Mobile IP becomes possible by adopting the new mobile IP protocol as it has correctly analyzed all the obstacles in standard mobile IP and proposes a complete solution fitting the different wireless technologies and the new business needs. Finally, the proposed correlation of the mobile users' credentials as IP, hostname, and network equipment identifiers with their (U)SIM cards provides a

 ISBN:978-1-61208-113-7

strong method for authenticating and authorizing mobile users while creating a unique universal identity that can reveal the real world identity. Without doubt this can facilitate the cyber crime investigation and enhance the cyber security.

REFERENCES

[1] A. Durresi1, L. E.s, V. Paruchuri, and L. Barolli: "Secure 3G User Authentication in Adhoc Serving Networks", Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06), June 2006.

[2] J. LaFlamme and M. Litwin, Deloitte Development LLC "Wireless and Mobility", 2010.

[3] H. Luo, P. Zerfos, J. Kong, S. Lu, and L. Zhang, "Self-securing Ad Hoc Wireless Networks", UCLA Computer Science Department.

[4] CAIMR (Center for applied Identity Management Research), "An Applied Research Agenda for confronting Global Identity Management Challenges, May 2009."

[5] C. Perkins, Ed. Nokia Research Center, "IP Mobility Support for IPv4," RFC 3344, August 2002.

[6] C. Perkins - Nokia Research Center, P. Calhoun - Cisco Systems, Inc., J. Bharatia - Nortel Networks, "Mobile IPv4 Challenge/Response Extensions (Revised)", IEFT RFC 3775, January 2007.

[7] G. Montenegro, Sun Microsystems, Inc., "Reverse Tunneling for Mobile IP", RFC 2344, May 1998

[8] F. Adrangi, Ed. Intel and H. Levkowetz, Ed. Ericsson, "Mobile IPv4 Traversal of Virtual Private Network (VPN) Gateways," RFC 4093, August 2005.

[9] Cisco Systems, Inc. , Design of the Mobile IP—Support for RFC 3519 NAT Traversal Feature, 2007

[10] Y. Rekhter - Cisco Systems, B. Moskowitz - Chrysler Corp., D. Karrenberg - RIPE NCC, G. J. de Groot - RIPE NCC, E. Lear - Silicon Graphics, Inc., "Address Allocation for Private Internets", RFC 1918, February 1996.

[11] S. Kent - BBN Corp, R. Atkinson - @Home Network, "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

[12] D. B. Johnson, C. E. Perkins and J. Arkko, "Mobility Support in IPv6" IEFT RFC 3775, June 2004.

[13] H.Soliman, C. Castelluccia, K. El-Malki, and L. Bellier, "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)-0," IETF RFC 4140, August 2005.

[14] W. F. Elsadek and M. N. Mikhail, Department of Computer Science and Engineering, the American University in Cairo "Universal Mobility with Global Identity (UMGI) Architecture", Proceedings 2009 International Conference on Wireless Networks and Information Systems (WNIS 2009), December 2009.

 ISBN:978-1-61208-113-7

A new Hybrid SPD-based Scheduling for EPONs

Qianjun Shuai, Jianzeng Li
Information Engineering Academy
Communication University of China
Beijing, China
{sqj, jzli}@cuc.edu.cn

Jinyao Yan, Weijia Zhu
Computer and Network Information Center
Communication University of China
Beijing, China
{jyan, wjzhu}@cuc.edu.cn

***Abstract*—Dynamic bandwidth allocation (DBA) is a key issue of Ethernet PONs. In order to get higher resource utilization and lower packet delay, the problem is always dissolved into grant sizing and grant scheduling. In this paper, we explore grant scheduling techniques. We propose a modified hybrid online and offline scheduling with the shortest propagation delay (SPD) first policy (we named HSPD) which can compensate for the idle time under light or medium loaded traffic. Meanwhile, the last ONU in offline set is adaptively indicated to transmit REPORT frame first (called LRF), so the idle time can be eliminated especially under heavy loaded traffic. We evaluate the cycle length and average packet delay through analysis and simulations. Compared with the offline SPD first scheduling (we named it OSPD), and online and offline scheduling with excess bandwidth distribution (so-called M-DBA1), we find out our algorithm HSPD-LRF can achieve significant improvements in terms of average packet delay and channel utilization.**

Keywords*—DBA; Online; Offline; Hybrid SPD-based Scheduling (HSPD); Last REPORT First (LRF).

I. INTRODUCTION

Passive optical network is a point-to-multipoint (P2MP) optical network without active elements in the path, where an optical line terminal (OLT) at the center office (CO) is connected to many optical network units (ONUs) at remote nodes through passive elements such as 1:N optical splitters. The network use a single wavelength in each of the two directions—downstream and upstream, and the wavelengths are multiplexed on the same fiber through coarse WDM (CWDM). In the downstream direction, packets are broadcast by the OLT and extracted by the destination ONU. While in the upstream direction all the ONUs share a single wavelength channel by time division multiple access (TDMA). Since the passive optical splitters are not able to inspect the upstream collision, it is the OLT who acts as the arbiter that grant timeslots (transmission windows) to the ONUs by a certain scheme. In order to avoid packets collision and to utilize the upstream channel efficiently, a dynamic bandwidth allocation (DBA) is required. Ethernet PONs (EPONs) technology has been standardized by the IEEE 802.3ah Ethernet in the First Mile (EFM) Task Force, which aims at combining the low-cost equipment, simplicity of Ethernet, the low-cost fiber infrastructure and high bandwidth of PONs. However IEEE 802.3ah does not prescribe any scheme of the upstream data transmission, but devises the multipoint control protocol (MPCP) which defines message-based mechanism to control information exchange between the OLT and ONUs. The shceme of upstream data transmission is left for choice of vendors. Dynamic bandwidth allocation (DBA) algorithms for EPONs have become a key issue and have been paid considerable attention from both industry and academia in recent years.

Basically, DBA algorithms that have been presented in literatures for EPONs can be divided into grant sizing and grant scheduling. The grant sizing determines the size of the upstream transmission window granted for an ONU. While the grant scheduling determines the beginning time of the upstream transmission granted for an ONU. These two aspects can be not separated. Grant sizing algorithms have been proposed in [5]-[9]. In [9], Kramer *et al.* proposed the IPACT algorithm in which the OLT polls the ONUs in a round-robin way and dynamically assigns them bandwidth according to different approaches and indicated that the limited service has the best performances. Many literatures proposed excess bandwidth allocation algorithms [2][5][6], or prediction mechanisms, based on the limited service [7]. Grant scheduling techniques have been proposed in [1]-[4]. The authors in [4] partition the scheduling problem into (1) a scheduling framework and (2) a scheduling policy operating within the adopted framework. McGarry *et al.* [3] outlines two basic

 ISBN:978-1-61208-113-7

grant scheduler as online and offline. In an online scheduler any ONU is scheduled for upstream transmission as soon as the OLT receives its REPORT message. In an offline scheduler the ONUs are scheduled for transmission of the next cycle once the OLT has received all REPORT messages from all ONUs. So online scheduling has great efficiency but lacks QoS control since the OLT makes scheduling decisions based on individual request without global knowledge of the current bandwidth requirements of the other ONUs. Offline scheduling allows OLT to take into consideration of the current bandwidth requirements from all ONUs, thus, it enables the wide variety of QoS mechanisms. On the other hand, it conducts idle time in upstream channel since OLT has to wait all REPORT frames from all ONUs.

In this paper, we aim at resolving the idle time issue of offline scheduling. In order to shorten or eliminate the fixed idle time of offline scheduling, we combine the online & offline scheduling based on limited service and sort the overloaded ONUs in ascending order by their propagation delays before scheduling in offline framework. Meanwhile in our proposed scheduling the last ONU is put into the offline scheduling set and the ONU with the largest propagation delay in the offline set transmits REPORT frame before its data in the next cycle. Since the ONUs in offline scheduling always have large enough data transmission window, the idle time can be compensated especially under heavy traffic. Importantly, the whole algorithm is very simple to implement.

The rest of this paper is organized as follows. In Section II, we discuss the related work on dynamic bandwidth allocation algorithms. In Section III, we analyze the idle time problem in offline scheduling and present the HSPD-LRF scheduling. In Section IV, we provide the performance result of simulations. In Section V, we conclude the paper.

II. RELATED WORK

Assi *et al*. [5] proposed an excessive bandwidth distribution that left from the underloaded ONUs amongst the overloaded ONUs (so-called DBA1). In order to implement the excessive bandwidth distribution, it for the first time employ a combined online and offline scheduling in which ONUs with requests smaller than its minimum guaranteed window sizes are scheduled immediately while those with larger requests would be scheduled when OLT have received all of the REPORT frames. Since DBA1 grants the total excess bandwidth to the overloaded ONUs, it sometimes mistakenly leaves most of the available bandwidth idle. Based on DBA1, Shami *et al*. in [6] proposed an improved dynamic bandwidth allocation algorithm called M-DBA1 which grants bandwidth to overloaded ONUs based on a comparison of total excess bandwidth saved by underloaded ONUs with total extra demand bandwidth of the overloaded ONUs. Bai *et al.* [8] improved the procedure for allocating excess bandwidth. Nevertheless, the algorithms mentioned above only focus on the grant sizing but not grant scheduling. They did not present any optimal scheduling policy but employed the simple first come first schedule (FCFS) scheme in the hybrid online and offline scheduler.

In [2], J. Zheng proposed a mechanism in which OLT always schedules underloaded ONUs before overloaded ONUs as well as possible. OLT maintains a time tracker to record the ending time of last scheduled ONU. When the upstream channel is going to be idle, an overloaded ONU is scheduled without extra excessive bandwidth if necessary. Thus, the upstream transmission channel is not idle between granting cycles.

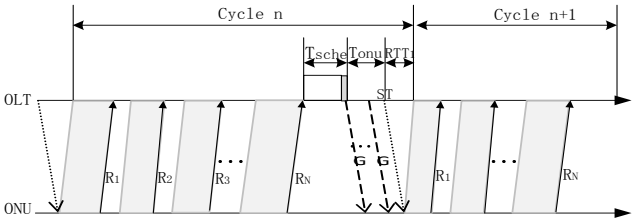
In [1], McGarry *et al*. presented a scheduling algorithm that employs the shortest propagation delay first (SPD) policy in offline framework. OLT sorts all ONUs in ascending order by their propagation delays before scheduling. Thus, the long round trip propagation delay can be masked by scheduling the near-by ONUs first.

III. HSPD-LRF ALGORITHM

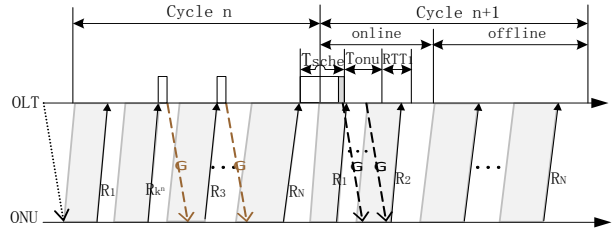
Since an offline scheduler makes scheduling decision for all ONUs at once, this requires that the scheduling algorithm be implemented after the OLT receives the end of the last ONU's REPORT frame. Thus, as illustrate in Figure 1(a), a fixed idle time between scheduling cycles is introduced. It is composite of the followings:

- The computation time of the scheduling in OLT T_{sche}.
- The transmission time for the grant (64 bytes) frame which can be regarded as part of scheduling time T_{sche}.
- The processing time of the ONU scheduled in the next

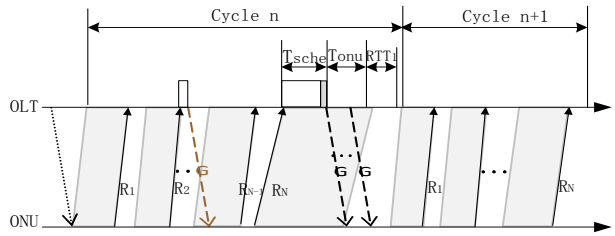
 ISBN:978-1-61208-113-7



(a) Offline Scheduling Framework



(b) Online and Offline (Hybrid) Framework



(c) LRF based Online and Offline Scheduling (Hybrid) Framework

Fig. 1 Different scheduling frameworks

cycle T_{onu}.

- The RTT of the first ONU scheduled in the next cycle RTT_1^{off}.

  So, we get:

$$T_{idle} = T_{sche} + T_{onu} + RTT_1^{off} - T_g \quad (1)$$

where T_g is the guard time between two consecutive transmission windows of two different ONUs. When the ONUs are scheduled based on SPD first policy, the RTT of the first ONU can be reduced to the minimum value. So, idle time in offline scheduler is shortened. Described as:

$$T_{idle}^{MIN} = T_{sche} + T_{onu} + RTT^{MIN} - T_g \quad (2)$$

However, the SPD based offline scheduling still has a fixed waste of idle time in upstream channel which reduces the channel utilization. In this paper, we employ a hybrid online and offline scheduling based on limited service and make changes in some aspects to address the idle time issue. First of all, all ONUs are divided into two sets according to their bandwidth requests which is given by (3). Those having bandwidth requests smaller than their minimum guaranteed windows are ascribed to light loaded set K and scheduled as soon as OLT receives their REPORT messages, the others are ascribed to over loaded set M and scheduled all at once according to their propagation delays sorted in ascending order after OLT receives the last REPORT message. Thus, the excessive bandwidth of underloaded ONUs can be used to meet the bandwidth demand of overloaded ONUs in each transmission cycle.

$$ONU_i \in \begin{cases} K, & R_i \le B_i^{MIN} \\ M, & R_i > B_i^{MIN} \end{cases} \quad (3)$$

As illustrated in Figure 1(b), if the first ONU who has request smaller that its minimum guaranteed window in the n*th* cycle is the k^n *th* to come. So, we get (4), which means the upstream channel idle time caused by the large RTT of the first ONU that belong to set K in the (n+1)*th* cycle and (5) which means the upstream channel idle time caused by the large RTT of the first ONU that belong to set M in the (n+1)*th* cycle. In the following, W_i is the transmission window length (time length) of the i*th* ONU in the n*th* cycle. R_j is the request length (time length) of the j*th* ONU. If both (4) and (5) are obtained simultaneously, we can achieve shorter idle time than the minimum idle time T_{idle}^{MIN} of the SPD based offline scheduling.

$$\sum_{i=k^n+1}^{N}\left(W_i + T_g\right) + T_{idle}^{MIN} + T_g \ge T_{sche} + T_{onu} + RTT_1^{on} \quad (4)$$

$$\sum_{j\in K^{n+1}}\left(R_j + T_g\right) + T_{idle}^{MIN} + T_g \ge T_{shce} + T_{onu} + RTT_1^{off} \quad (5)$$

Here, we get:

$$\sum_{i=k^n+1}^{N}\left(W_i + T_g\right) + RTT^{MIN} \ge RTT_1^{on} \quad (6)$$

$$\sum_{j\in K^{n+1}}\left(R_j + T_g\right) + RTT^{MIN} \ge RTT_1^{off} \quad (7)$$

A specific situation is when $k^n = N$, that means only the

 ISBN:978-1-61208-113-7

last ONU (the N*th* ONU) has smaller request, so, the transmission window ($W_i + T_g$) in (6) is nothing. Thus, only when $RTT_1^{on} = RTT^{MIN}$ could we get no longer idle time than T_{idle}^{MIN} if the last ONU is scheduled in online set. Due to this, the last ONU is always put into offline set. That means $ONU_N \in M$. Also we can see from (7) that the more elements in set K, the better.

When the traffic is getting heavier, more and more ONUs belonged to overloaded set M. So, the underloaded ONUs may not be able to compensate the idle time. In this situation, the last ONU of set M is indicated to transmit REPORT frame before its data, see Fig. 1(c). Since the ONUs in set M always has long enough data transmission window, it can always compensate the idle time. Specifically, this is easily executed by redefining the MPCP GATE frame granting to the ONUs. It is assumed that the general GATE frame can offer no more than 3 grants to an ONU, which means the valid range of the GATE number is from 0~3. As shown in Figure 2, the combining of the most significant bit and the third bit from the right of the Number/Flag byte in GATE frame can be defined as the indication of transmitting REPORT first.

IV. PERFORMANCE EVALUATION

To evaluate the performances such as the average cycle time, the average packet delay and channel utilization of the proposed LRF-based hybrid scheduling which we call HSPD-LRF, a simulation model comprising an access network with one OLT, and 16 ONUs was developed using C++. Here, channel utilization was defined as the ratio of the sum of pure data transmission windows to the cycle time. The pure data transmission window did not include the overheads, such as Preamble, IPG and the REPORT frame that attached. In the simulation, the equal weighted limited grant sizing with excess bandwidth distribution [6] is employed. In addition, the traffic of each ONU was generated with the properties of self-similarity and long-range dependence, and the Hurst parameter was set to 0.8. The maximum cycle time was assumed to be 2ms. The guard time between two consecutive transmission windows of two different ONUs was 1μs. The corresponding minimum guaranteed window size, B_i^{MIN}, was set

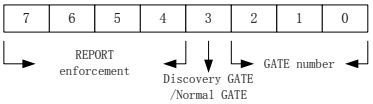


Fig. 2 Number/Flag byte in MPCP GATE

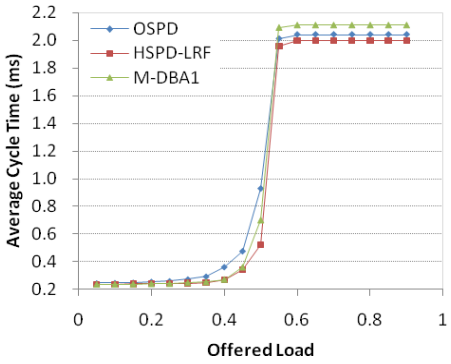


Fig. 3 Average cycle time versus traffic load

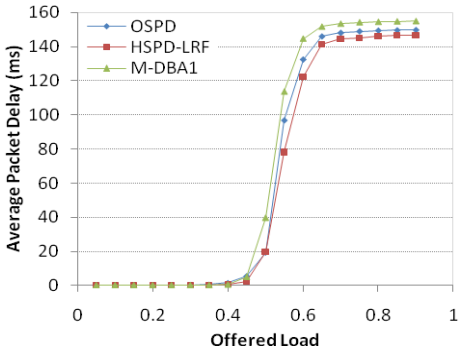


Fig. 4 Average packet delay versus traffic load

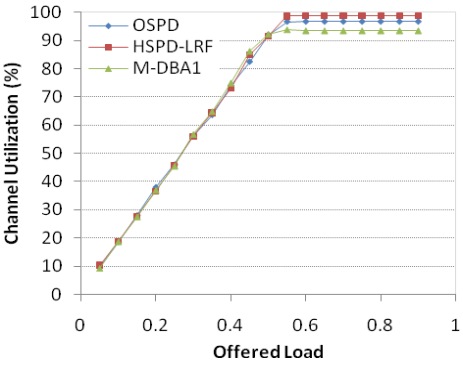


Fig. 5 Channel utilization versus traffic load

to15500 bytes. In order to explicit the impact of the diversity of propagation delays, the one-way propagation delays between the ONUs and the OLT were randomly generated according to a uniform distribution with a minimum value of

10μs and a maximum value of 200μs to represent the distance between 1km and 20km. Among others, one ONU had the minimum propagation delay, 10μs, and another had the maximum propagation delay, 200μs. The performances of the offline SPD based scheduling (OSPD) and the so-called M-DBA1 algorithm were also illustrated as comparison.

Figure 3 shows that the overall cycle time of the proposed algorithm is shortened compared with OSPD scheduling. When the traffic load is getting heavier (load > 0.4), HSPD-LRF has the shortest cycle time because it efficiently eliminates the idle time. Figure 4 shows the average packet delay versus offered load. Again the proposed algorithm has the best performance for all traffic load. Figure 5 is the channel utilization. Under heavy traffic load, the proposed algorithm achievs about 98.6% channel utilization which actually means there is no idle time wasted because the channel utilization was calculated using the pure data transmission windows without the overheads and the REPORT frames, which in fact occupy the upstream channel.

V. CONCLUSION AND FUTURE WORK

This study has presented a modified hybrid online and offline scheduling algorithm for EPONs, in which the underloaded ONUs are scheduled instantaneously without any delay, and the idle time issue is solved by adaptively employing the last REPORT first scheme when the traffic load is getting heavier. Specifically, the whole algorithm is executed simply since it only needs to sort the overloaded ONUs once before they are scheduled in offline mode as well as the OLT indicate the last ONU in overloaded set to transmit REPORT before its data based on slightly modified MPCP GATE frame when necessary. Through simulation results, the proposed algorithm has demonstrated that it can significantly improve the network performance in terms of packet delay and channel utilization as compared with the SPD-based offline scheduling and the well known M-DBA1 algorithm proposed in [6]. However, this study only investigated the network performances of the single channel EPON system. In the future work, the authors will investigate the issues in the multi-channel EPON, such as WDM EPON system.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of P. R. China (No. 60970127), Science Technology Research key Project of Ministry of Education (No.109029) and partly supported by Program for New Century Excellent Talents in University (NCET-09-0709). It is also supported by the Campus Engineering Project (No. XNG0943).

REFERENCES

[1] M. McGarry, M. Reisslein, F. Aurzada, and M. Scheutzow, "Shortest Propagation Delay (SPD) First Scheduling for EPONs with Heterogeneous Propagation Delays," Selected Areas in Communications, IEEE Journal on Volume: 28 , Issue: 6 , Page(s): 849 – 862,2010.

[2] J. Zheng, "Efficient bandwidth allocation algorithm for Ethernet passive optical networks," IEE Proc.-Commun., vol. 153, no. 3, pp. 464–468, June 2006.

[3] M. McGarry, M. Maier, and M. Reisslein, "WDM Ethernet Passive Optical Networks," *IEEE Communications Magazine*, vol. 44, no. 2, pp. S18–S25, February 2006.

[4] M. McGarry *et al.*, "Just-in-time scheduling for multichannel EPONs," *J. Lightwave Technol.*, vol. 26, no. 10, pp.1204–1216, May 2008.

[5] C. Assi, Y. Ye, S. Dixit, and M. Ali, "Dynamic Bandwidth Allocation for Quality-of-Service over Ethernet PONs," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 9, pp. 1467–1477, November 2003.

[6] A. Shami, X. Bai, C. Assi, and N. Ghani, "Quality of Service in Two-Stage Ethernet Passive Optical Access Networks," in Proceedings of IEEE the 13th International Conference on Computer Communications and Networks (ICCCN), 2004, pp. 352–357 Chicago.

[7] Y. Luo and N. Ansari, "Limited Sharing with Traffic Prediction for Dynamic Bandwidth Allocation and QoS Provisioning over Ethernet Passive Optical Networks," *OSA J. Opt. Net.*, vol. 4, no. 9, Sept. 2005, pp. 561–72.

[8] X. Bai, A. Shami and C. Assi, "On the fairness of dynamic bandwidth allocation schemes in Ethernet passive optical networks," *Journal of Computer Communications*, vol. 29, no. 11, pp. 2123-2135, July 2006.

[9] G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT: A dynamic protocol for an Ethernet PON (EPON)," *IEEE Communications Magazine*, 40(2):74–80, February 2002.

 ISBN:978-1-61208-113-7

Link Emulation on the Data Link Layer in a Linux-based Future Internet Testbed Environment

Martin Becke, Thomas Dreibholz, Erwin P. Rathgeb
University of Duisburg-Essen
Institute for Experimental Mathematics
Ellernstrasse 29, 45326 Essen, Germany
{martin.becke,dreibh,rathgeb}@iem.uni-due.de

Johannes Formann
University of Duisburg-Essen
Institute for Computer Science
Schützenbahn 70, 45117 Essen, Germany
jformann@dc.uni-due.de

***Abstract*—Protocol design and development is not a straightforward process. Each approach must be validated for interactions and side-effects in the existing network environments. But the Internet itself is not a good test environment, since its components are not controllable and certain problem situations (like congestion or error conditions) are difficult to reproduce. Various testbeds have been built up to fill this gap. Most of these testbeds also support link emulation, i.e. using software to mimic the characteristic behaviour of certain kinds of network links (like bandwidth bottlenecks or error-prone radio transmissions). The most popular link emulation systems are the Linux-based NETEM and DUMMYNET, which are e.g. applied on the IP layer of Planet-Lab and various other testbeds. However, the restriction to the OSI Network Layer (here: IP) is insufficient to test new non-IP Future Internet protocols.**

In this paper, we first introduce DUMMYNET and NETEM. After that, we will present our approach of adapting DUMMYNET for Linux to support link emulation on the Data Link Layer. Finally, we evaluate the applicability and performance of DUMMYNET and NETEM for link emulation on the Data Link Layer, in a Planet-Lab-based testbed environment. Our goal is to outline the performance and limitations of both approaches in the context of Planet-Lab-based testbeds, in order to make them applicable for the evaluation of non-IP Future Internet protocols.[1]

Keywords: Link Emulation, Data Link Layer, Future Internet Testbed, NETEM, DUMMYNET

I. INTRODUCTION

The protocol development for the current Internet bases on a strictly hierarchical structure, which has been standardized as the OSI reference model [1]. This model covers classic Internet applications like e-mail and file transfer quite well. However, the rapid technological developments driven by the needs of new applications (e.g. mobility, e-commerce, VoIP) show the conceptual limitations of this approach. Solutions like cross-layer optimization weaken the hierarchical structure and make the resulting protocol implementations difficult to develop and maintain. This makes the realization and deployment of new features and protocols, e.g. multipath TCP [2], difficult. On the other hand, clean-slate service-oriented frameworks try to solve such conceptual issues by completely getting rid of the hierarchical structure. Multiple large research projects like FIRE and G-Lab in Europe, GENI and FIND in the U.S.A. and AKARI in Asia examine such approaches.

[1]Parts of this work have been funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) and the German Research Foundation (Deutsche Forschungsgemeinschaft)

In every case, protocol design and development is more than a theoretical procedure. Over the years, protocol and network developers strike new paths to validate their approaches and concepts. Long-term experience shows that protocol development involves several more steps besides requirement formulation, specification, verification and documentation. In particular, it also includes the test of implementation and conformance as well as analysing and optimizing the performance for special applications and architectures. This requires the integration of the new ideas and approaches into real hardware, emulation and also into simulation testbed environments. That is, each new approach and proof of concept needs an adequate testbed. But each research community has its own specific requirements on its testbeds, resulting in many different variants. Examples of such testbeds are Emulab [3], VINI [4], One-Lab [5], G-Lab [6] and Planet-Lab [7].

Link emulation (i.e. using software to e.g. apply delay or bandwidth limitations of a satellite link to certain packet flows) is a widely used feature in such testbeds. Currently, the testbeds apply this link emulation feature on IP flows (i.e. on the Network Layer). While this is sufficient within IP networks – i.e. the current Internet protocols – it prevents the usage for new non-IP Future Internet approaches. But if projects are investigate on clean clean-slate service-oriented framework below IP – like German Lab project – emulation on a layer below is needed. Therefore, a link emulation solution on the Data Link Layer is desirable within testbeds. DUMMYNET [8], [9] (e.g. used by Planet-Lab, G-Lab and Emulab, see [10]) and NETEM [11] are popular tools for link emulation on Linux, which is the operating system used by the majority of the testbeds. NETEM already supports Data Link Layer usage, while DUMMYNET under Linux had lacked of this feature. In this paper, we first introduce our approach of extending the Linux version of DUMMYNET by Data Link Layer support. Then, we evaluate the performance of both approaches in a Planet-Lab-based testbed environment – used in the German Lab project – which we have extended by the support of link emulation on the Data Link Layer. The challenge here is to compare at first time link emulation an data link layer on one operation system.

 ISBN:978-1-61208-113-7

II. Link Emulation

The intention of testbeds is to examine concepts and architectures under different conditions, e.g. nearly optimal conditions for a first proof of concept or more realistic conditions to analyse specific scenarios like communication over modem, satellite or other wireless links.

A. Constituting Physical Constraints in a Testbed

Clearly, the physical model (e.g. the underlying transmission technology) is an important part of such tests, because it realizes the constraints set by physics and hardware implementations. The most realistic results on testing a system with certain hardware constraints is to actually run it on real hardware. This approach is e.g. used by the G-Lab testbed [6]. However, real network behaviour – like the effects caused by background traffic or BGP routing in the Internet – are not easily reproducible in such hardware-centric systems. Approaches like the Planet-Lab [7] interconnect a large number of virtualized systems over the real Internet, allowing for overlay network tests. However, in this case, the hardware itself is not under the control of the researcher.

For many upper-layer test cases, an emulation model of the physical constraints is already sufficient, e.g. a satellite link with its typical delay and bit errors may be emulated in software. Such an emulation is also easily adaptable to special cases and allows for easy reproduction of the results, e.g. to examine the communications protocol performance over a satellite link during solar wind. Particularly, software emulation is inexpensive, since special hardware (e.g. a real satellite link) are not needed. The most well-established emulation systems in the context of testbeds emulation systems are DUMMYNET [9] and NETEM [11]. These tools emulate links with variable delay, bandwidth, packet loss and jitter, i.e. they can be used to model the link characteristics of different network access technologies. We will introduce DUMMYNET and NETEM in the following subsections.

B. NETEM

The Linux Advanced Traffic Control Framework [12], which is part of the Linux kernel, uses filtering rules to map packets or frames – i.e. data on Data Link as well as Network Layers – to queuing disciplines (QDisc) of an egress network interface. QDiscs may be classful, i.e. contain a hierarchy of subclasses. Filtering rules (denoted as classifier) of the QDisc itself map packets or frames to the subclasses. Each subclass may have its own QDisc, which again may be classful or classless (i.e. no subclasses). NETEM [11] is a classful QDisc for Linux. This QDisc itself provides packet delay, loss, duplication and re-ordering.

If bandwidth limitations are required, a secondary queuing discipline – like the classless Token Bucket Filter (TBF) QDisc – has to be applied as sub-QDisc of NETEM to control the data rate. Bandwidth limitations are always based on the Data Link Layer frame sizes of the egress interface on which NETEM and subservient QDiscs are configured on. That is, using e.g. NETEM and TBF on an Ethernet interface, all bandwidth calculations for packets include the Ethernet headers and trailers.

C. DUMMYNET

DUMMYNET [9] provides link emulator functionality in the FreeBSD kernel. The packet filtering architecture of the kernel is used to pass packets through one or more queues. A hierarchy of the queues is realized by so-called pipes. A pipe represents a fixed-bandwidth channel; queues actually store the packets. Each queue is associated with a weight. Proportionally to its weight, it shares the bandwidth of the pipe it is connected to.

Originally, DUMMYNET had been realized on the Network Layer to control bandwidth, delay and jitter as well as packet loss rate, duplication and reordering of IP packet flows. A recent patch [13] for FreeBSD has added support for the Data Link Layer, i.e. the patched DUMMYNET implementation is able to handle frames on the Data Link Layer containing arbitrary Network Layer traffic. This allows for applying DUMMYNET to handle non-IP Future Internet protocols. [8] introduces a port of DUMMYNET to Linux, in order to apply it for Planet-Lab-based G-Lab Experimental Facility. However, this port does *not* support Data Link Layer traffic, due to the significantly different handling procedures of Data Link Layer frames in Linux. It is in the end an adaptation of a new packet filter mechanism on the Data Link Layer.

Bandwidth limitations realized by DUMMYNET always base on the Network Layer packet size only, even if DUMMYNET is used to restrict Data Link Layer traffic. That is, using DUMMYNET e.g. to shape IP traffic over an IEEE 802.11 WLAN does *not* include the WLAN headers and trailers.

III. DUMMYNET ON LINUX

The link emulation infrastructure of Planet-Lab-based G-Lab – as well as of its derived testbeds – is based on the Linux port of DUMMYNET [8]. In order to extend the G-Lab infrastructure by DUMMYNET-based Data Link Layer support, we had to extend DUMMYNET for Linux.

Figure 1 presents our concept for extending DUMMYNET on Linux by the support of link emulation on the Data Link Layer. The existing DUMMYNET hooks into the Network Layer packet chains of the Routing Subsystem. Chains are used by the packet filtering architecture of Linux and provide mechanisms to intercept and manipulate packets. Before a packet is routed, it traverses the PREROUTING chain. Packets to be forwarded to another system then pass through the FORWARD chain while packets destined for the system itself are handled by the INPUT chain. Packets sent from the system itself come from the OUTPUT chain. After routing, a packet passes through the POSTROUTING chain. The two hooks used by DUMMYNET are on the PREROUTING and POSTROUTING chains. DUMMYNET can intercept a packet, and eventually return it some time later into its original chain.

This Network Layer concept for DUMMYNET can be extended to the Data Link Layer in the Bridging Subsystem. Note, that Linux uses the same chain naming (i.e. PREROUTING, FORWARD, INPUT, OUTPUT, POSTROUTING) as for the Routing Subsystem. Nevertheless, the Routing and Bridging Subsystems are completely independent. DUMMYNET has to be extended by the support for hooking also into the

 ISBN:978-1-61208-113-7

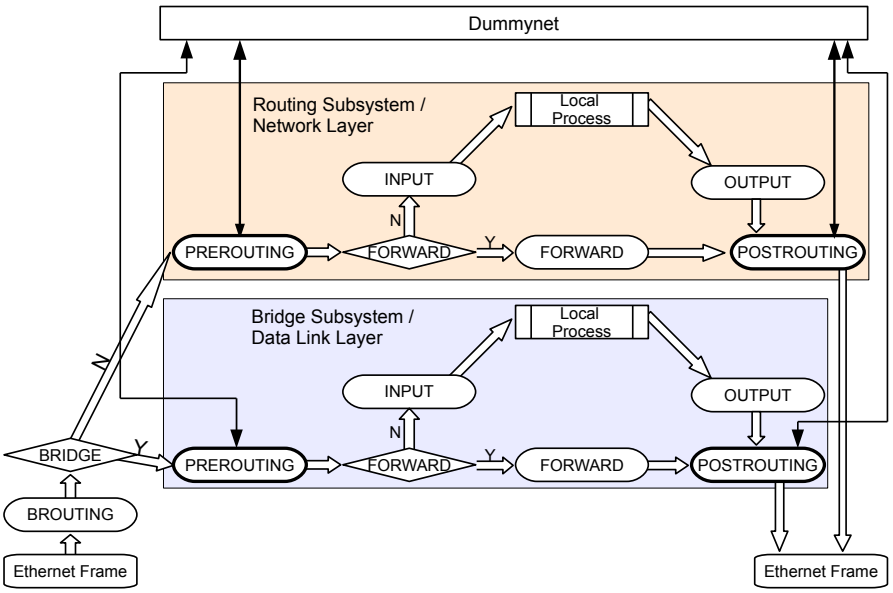


Figure 1. Our Concept of Extending DUMMYNET on Linux

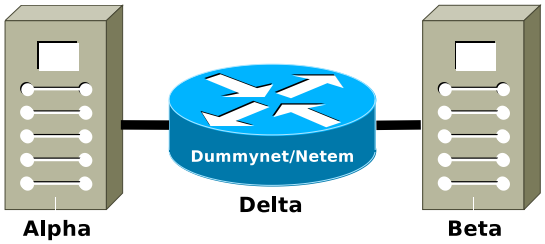


Figure 2. Testbed Setup

chains of the Bridging Subsystem to intercept frames on the Data Link Layer. Furthermore, the internal data handling of DUMMYNET has to be adapted to handle the frame structures.

The implementation of our approach has been realized as a patch for the Linux kernel, supporting all functionalities being necessary for our performance evaluation. After extending additional filtering features being necessary for real G-Lab deployment, we will contribute our patch to the G-Lab kernel development process. In this context it should be noticed, that this changes are independet from the infrastructure of planetlab, so that there should no drawbacks for this approach.

IV. TESTBED ENVIRONMENT

In order to evaluate the performance of DUMMYNET and NETEM, we have set up a G-Lab-based Linux test environment as shown in figure 2: the sender node "Alpha" is connected via router "Delta" to the receiver node "Beta" over 100 Mbit/s full-duplex Ethernet links. Table I provides the technical details of the systems; their hardware has been intentionally chosen to be "low performance", in order to illustrate the effects of high CPU load on the DUMMYNET/NETEM performance (which is a likely situation for G-Lab nodes hosting a large number of active slices). Router "Delta" is configured with both, DUMMYNET (including our Data Link Layer support extension as described in section III) and NETEM. The Planet-Lab-based test infrastructure has been extended to support link emulation on the Data Link Layer in addition to the already existing link emulation on the Network Layer. A configuration option decides whether to apply DUMMYNET or NETEM.

NUTTCP [14] has been used for throughput and packet loss measurements using the UDP protocol. Due to the different traffic measurement bases of NETEM (Data Link Layer frame size, see subsection II-B) and DUMMYNET (Network Layer packet size, see subsection II-C), we have configured the packet output rate R^{nuttcp} of NUTTCP appropriately to achieve a desired on-network Data Link Layer rate R_{Network} for a given payload message size M and IP header size H_{IP}, UDP header size H_{UDP} and Ethernet header/trailer size H_{Ethernet}:

$$R^{\text{nuttcp}}_{\text{NetEm}} = \frac{M * R_{\text{Network}}}{M + H_{\text{IP}} + H_{\text{UDP}}} \quad (1)$$

$$R^{\text{nuttcp}}_{\text{Dummynet}} = \frac{M * R_{\text{Network}}}{M + H_{\text{IP}} + H_{\text{UDP}} + H_{\text{Ethernet}}} \quad (2)$$

Our bandwidth results always show the achieved Data Link Layer throughput at the receiver side.

For measuring delay, we have utilised the standard PING tool (which uses ICMP Echo Requests and Replies).

V. PERFORMANCE ANALYSIS

In the following analysis, we examine the performance of DUMMYNET and NETEM based on the studies in [8], [11], [15]. But unlike former studies, our interest is in the performance of link emulation on the Data Link Layer, which has not been examined before – but which becomes crucial when examining Future Internet protocols on top of it. In the following subsections, we evaluate Planet-Lab-based setups

 ISBN:978-1-61208-113-7

Node Name	Processor	Memory	Role
Alpha	700 MHz AMD Duron	512 MiB	Sender
Beta	700 MHz AMD Duron	512 MiB	Receiver
Delta	1666 MHz AMD Athlon	1024 MiB	Router

Table I
TECHNICAL DETAILS OF OUR G-LAB-BASED TESTBED

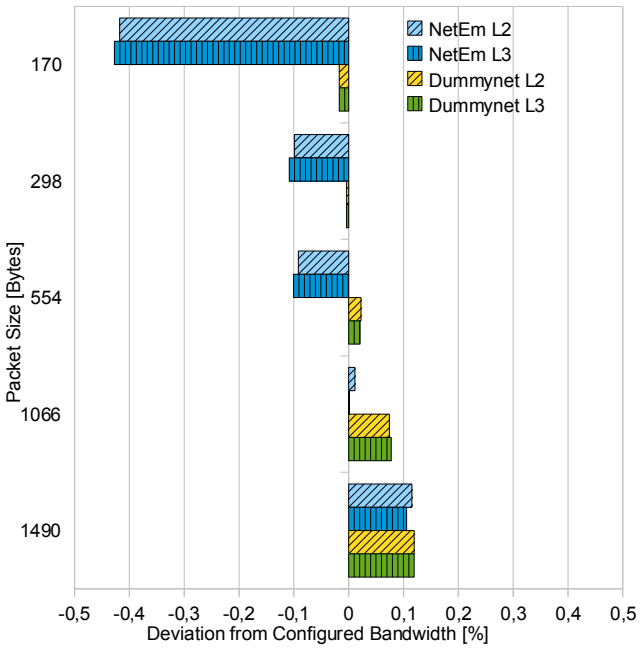


Figure 3. Derivation from Configured Bandwidth

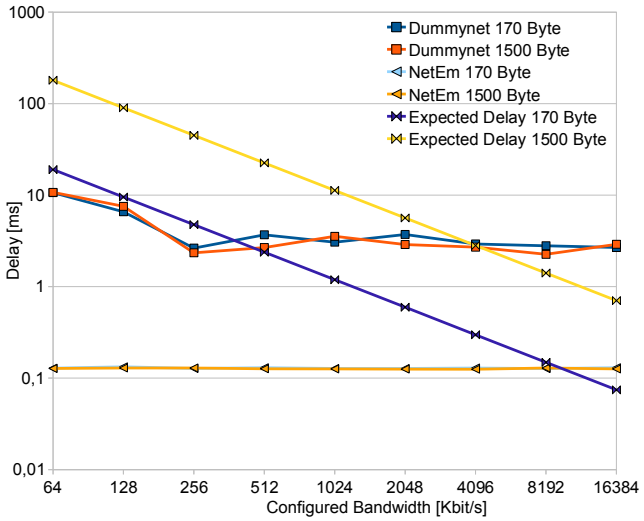


Figure 4. Expected and Measured Delays for Different Packet Sizes

with the link emulation varying the basic QoS measures: error rate, bandwidth limitation, delay and jitter.

A. Bandwidth

Bandwidth is probably the most crucial QoS measure. Therefore, the accurate adherence of configured bandwidth limitations – for any kind of traffic pattern – is a highly important feature of a link emulation system. Figure 3 shows the deviations from the desired bandwidth of 8389 Kbit/s (setting based on a DSL media streaming scenario) for DUMMYNET and NETEM on Data Link (L2 – Layer 2) and Network (L3 – Layer 3) Layers when varying the packet size. Small packets particularly occur in multimedia scenarios, leading to a high per-byte routing/bridging overhead. DUMMYNET and NETEM handle different packet sizes on both layers quite well, with an increased deviation for NETEM when using small packets (about -0.4% for 170 byte packets). However, this deviation still remains small and should be uncritical for most use cases.

In this context, it has to be mentioned that tests configured with relatively short traffic duration time and large buffers – over particularly low-bandwidth emulated links – can lead to distortions of the measurements by the time necessary to fully transmit the buffered packets. Furthermore, due to the bandwidth-delay product, packets with larger size require a proportionally longer transmission time on the link. This delay – depending on the configured bandwidth and used packet sizes – is *not* being emulated by DUMMYNET or NETEM. An example is provided in figure 4, which shows the expected and measured delays for varying bandwidth using packets of 146 bytes and 1494 bytes. The user should be aware of this fact if packet sizes differ significantly.

In case of bandwidth limitation, it is important to figure out the limitation of the testbed hardware. Depending on the needs of the applications, the traffic pattern could differ in the size of the messages, e.g. small messages in a multimedia setup or full MTUs in a download scenario. The resulting loss rate for enforcing a maximum bandwidth of 100 Mbit/s using packet sizes of 146 bytes and 1494 bytes is shown in figure 5; the left-hand plot presents the DUMMYNET results, the right-hand plot the NETEM results. Since the data rate generated by NUTTCP is less or equal to the enforced data rate, no loss should occur. But obviously, small packet sizes result in packet losses [16], since the hardware (CPU, but also network interface cards and buffers) are incapable of handling the high number of packets per second. This side effect – which is caused by interrupt frequency, timer resolution, the latency of context-switch operations by the operating system and also limited queue sizes [17] – has to be considered carefully when planning experiments in the testbed. These limitations apply for both, Data Link as well as Network Layer link emulation. But in comparison to NETEM, DUMMYNET shows these side effects earlier: at a rate of about 30 Mbit/s and a packet rate

 ISBN:978-1-61208-113-7

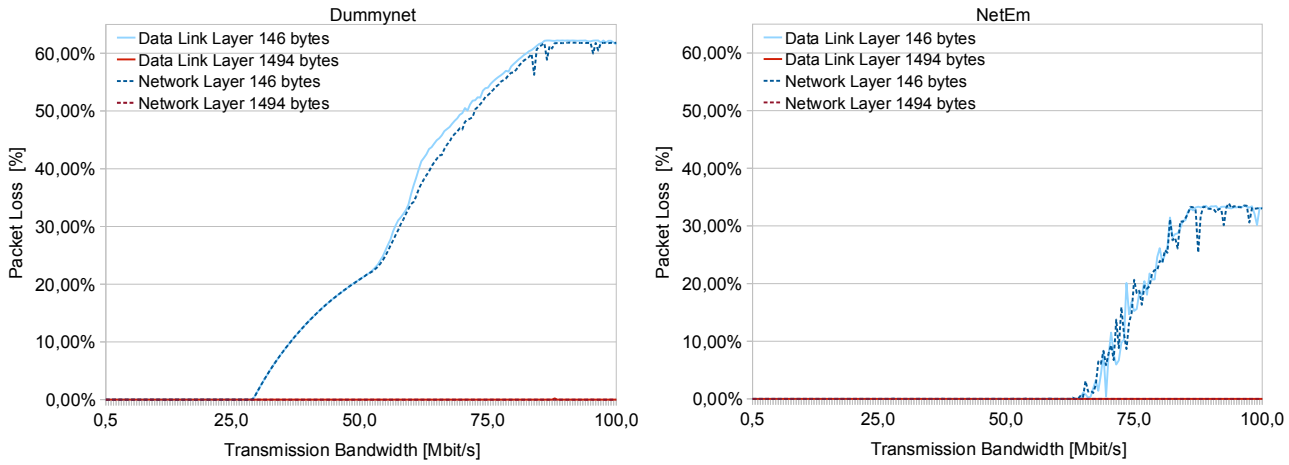


Figure 5. Hardware Limitations of DUMMYNET and NETEM for Different Packet Sizes

of 146 bytes, DUMMYNET starts losing packets while NETEM is capable of handling more than 60 Mbit/s without losses.

B. Error Rate

The loss rate is another important QoS measure for link emulation, e.g. to emulate lossy wireless or satellite connections. Therefore, it is useful to examine the accuracy of DUMMYNET and NETEM to adhere to a configured loss rate. For our measurement, NUTTCP has generated the configured data rate, while the link emulation has applied a configured loss rate. Figure 6 shows the absolute deviation from the configured loss rates for DUMMYNET and NETEM, using Data Link Layer (Layer 2) as well as Network Layer (Layer 3) link emulation. Since small packets are the most performance-critical (due to the high rate of packets/s), we only present the results for a packet size of 146 bytes here.

While the deviations for a NUTTCP bandwidth of 1 Mbit/s remain small for both, DUMMYNET and NETEM, a significant deviation is observable already for DUMMYNET at 40 Mbit/s. Even higher deviations can be found at a bandwidth of 70 Mbit/s. In this case, also NETEM shows an increased loss rate deviation, but this is still significantly smaller than for DUMMYNET. The reason of this deviation for DUMMYNET is again the limitation of the system resources, as already observed for the bandwidth limitation in subsection V-A. Again, NETEM can cope significantly better with these limited resources and still achieve a reasonable performance in parameter ranges where DUMMYNET is unable to work properly any more. This property applies for Data Link as well as Network Layer link emulation.

C. Delay

In order to examine the accuracy of the link emulation to adhere to a configured packet delay, we have varied the desired delay in a 100 Mbit/s setup for packet sizes of 170 bytes and 1500 bytes (i.e. full MTU on the Network Layer). The results are presented in table II for DUMMYNET and NETEM on Data Link (L2 – Layer 2) and Network (L3 – Layer 3) Layers. For both layers, the differences between the two packet sizes (i.e. small packets vs. full MTU) are quite small. However, it is observable that the delay results achieved by NETEM more accurately reach the configured target delay. For example, the difference to a target delay of 100 ms is almost 2 ms for DUMMYNET, but only 0.01 ms for NETEM. Also, the delay achieved by DUMMYNET is a little bit smaller than the actual target delay in most cases. This may distort measurements expecting a hard lower bound on the packet latency.

D. Jitter

While NETEM provides a configuration option to apply certain jitter distributions to the traffic, jitter is not directly supported by DUMMYNET. In DUMMYNET, jitter can be mimicked by configuring a set of pipes with different delays. Traffic is mapped to these pipes appropriately to reach a certain delay distribution. Due to these differences, it is not possible to directly compare the jitter performance of both approaches. We therefore show the Data Link and Network Layer performances of both systems separately.

For our jitter examination, we have configured a 100 Mbit/s setup using 1500 byte packets (i.e. full MTU). NETEM has been configured with a normal delay distribution of 100 ms average, while DUMMYNET has been set up with 11 pipes to mimic a similar distribution. The delay distributions are presented in figure 7; the left-hand plot shows the DUMMYNET results, the right-hand plot the distribution for NETEM. The Network Layer (Layer 3) values are displayed by the bars, the line depicts the Data Link Layer (Layer 2) values. As expected, the behaviour for Data Link and Network Layers is quite similar. Due to the different capabilities of NETEM, the distribution for NETEM is quite smooth while for DUMMYNET the 11 pipes are clearly observable as large peaks. To achieve a smoother distribution, DUMMYNET could be set up with a larger number of pipes. However, the complexity and resource consumption of such a kind of configuration would be extraordinarily high.

 ISBN:978-1-61208-113-7

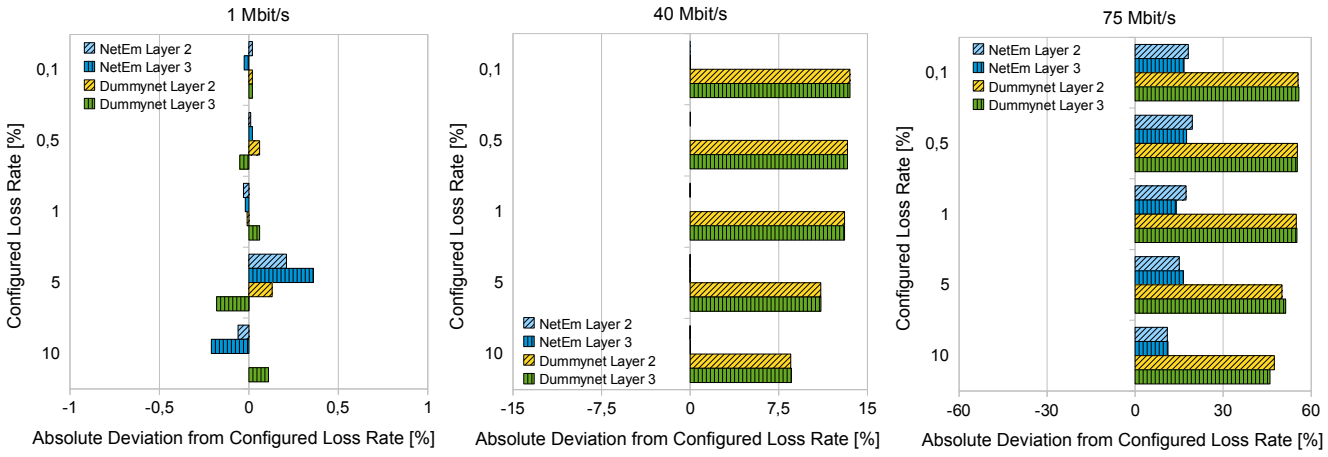


Figure 6. Derivation from Configured Error Rate for using DUMMYNET and NETEM

	Delay in ms for packet size of 1500 Byte							
Emulation\Parameter	5	10	20	50	100	200	500	1000
Dummynet L2	5.9	10.5	18.3	50	98.3	198.1	498.1	998.1
Dummynet L3	5.9	10.5	18.3	50.2	98.4	198.2	498,1	998.2
NetEm L2	5	10	20	50	100	200,1	500	1000
NetEm L3	5	10	20	50	100	200.2	500.1	1000.1

	Delay in ms for packet size of 170 Byte							
Emulation\Parameter	5	10	20	50	100	200	500	1000
Dummynet L2	5,97	11.69	18.4	50.84	98.24	198.4	498.25	998.35
Dummynet L3	5,93	11.69	18.62	50.54	98.2	198.41	498.26	998.22
NetEm L2	5	10	20.01	50	100.01	200.01	500.01	1000.01
NetEm L3	5	10	20.01	50.01	100.01	200.01	500.01	1000.01

Table II
DELAY ON DATA LINK AND NETWORK LAYER

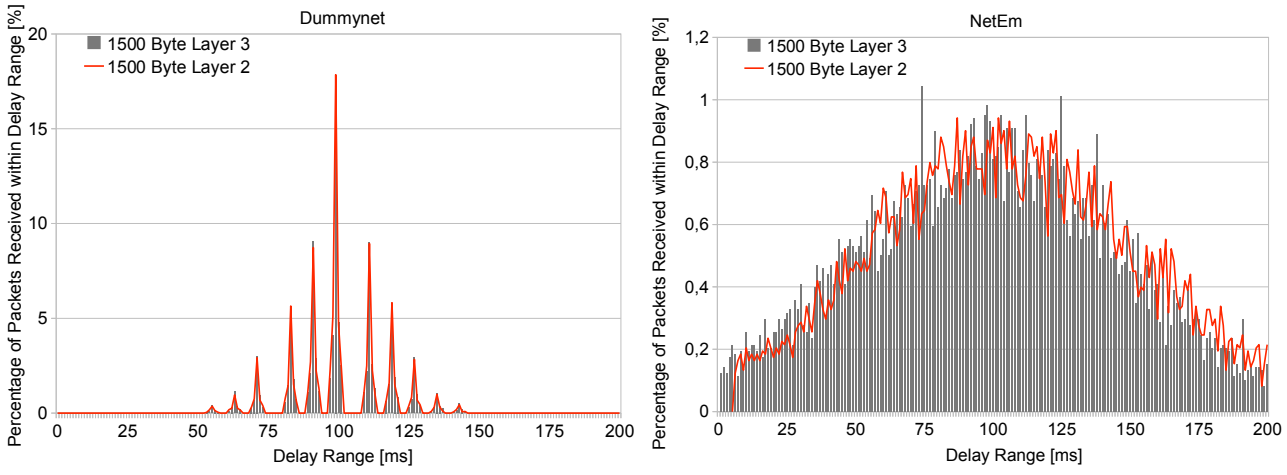


Figure 7. Emulating Jitter with DUMMYNET and NETEM

	146 Byte Packets					1500 Byte Packets	
Bandwidth	Baseline	L2 NETEM	L2 DMYNET	L3 NETEM	L3 DMYNET	L2 DMYNET	L3 DMYNET
1 Mbit/s	0.03%	0.04%	0.96%	0.04%	0.96%	0.10%	0.10%
15 Mbit/s	0.28%	0.39%	13.83%	0.50%	14.51%	1.56%	1.52%
30 Mbit/s	0.75%	0.88%	28.26%	1.01%	30.02%	2.93%	3.07%
45 Mbit/s	1.12%	1.71%	52.44%	1.58%	54.33%	4.46%	4.56%
60 Mbit/s	1.78%	2.15%	60.72%	2.18%	62.97%	5.18%	5.29%
75 Mbit/s	1.57%	2.15%	69.74%	2.19%	73.74%	5.71%	6.02%

Table III
CPU UTILIZATION FOR USING DUMMYNET AND NETEM WITH DIFFERENT PACKET SIZES

E. CPU Load

Beside the four network QoS measures, it is also important to examine the CPU load caused by the link emulation. In particular, Planet-Lab nodes are usually highly loaded by their slices already. The link emulation should therefore save CPU resources and – even more important – not exceed the CPU capacity (which would cause undesired frame/packet loss).

Table III shows the CPU load caused by DUMMYNET and NETEM for applying bandwidth limitation on Data Link (L2 – Layer 2) and Network (L3 – Layer 3) Layers for packet sizes of 146 bytes and 1500 bytes (i.e. full MTU). The link emulation is just used to enforce the configured bandwidth by dropping out-of-profile packets. For comparison, also a baseline measurement for 146 bytes packets (i.e. the performance-critical case) without link emulation is shown. Clearly, without link emulation, the CPU utilization remains quite small: about 1.5% for a 75 Mbit/s bandwidth limitation regardless of the packet size. Also, using NETEM on Data Link or Network Layer only slightly increases the CPU utilization to about 2.2% – regardless of the packet sizes. On the other hand, the CPU load for DUMMYNET is significantly influenced by the packet size: for 1500 byte packets, it requires 5.7% (Data Link Layer) and 6.0% (Network Layer) of the CPU, while the load rises to 69.74% (Data Link Layer) and 73.74% (Network Layer) for the small 146 bytes packets. That is, DUMMYNET requires significantly more CPU power for the same task.

VI. CONCLUSIONS

In this paper, we have evaluated our approach of extending Planet-Lab-based network testbeds – with fokus on G-Lab – by emulation on the Data Link Layer of the OSI model. Unlike the already existing link emulation supported on the Network Layer (i.e. for the IP protocol) only, our approach also allows for testing new non-IP Future Internet protocols. Two popular link emulation approaches have been considered: DUMMYNET and NETEM. The DUMMYNET approach – which is currently used by Planet-Lab for Network Layer link emulation – first had to be extended by us to support link emulation on the Data Link Layer of Linux-based Planet-Lab setups.

In our evaluation, we have shown that both Data Link Layer link emulation approaches are usable for Planet-Lab. The resulting performance of NETEM and DUMMYNET for the Data Link Layer emulation is quite similar to the performance of the Network Layer emulation. However, NETEM provides a slightly better accuracy for delay emulation and requires significantly less CPU power in comparison to DUMMYNET. Also, NETEM is able to emulate jitter much more accurately. As part of our future work, we are therefore going to contribute a configurable link emulation solution to Planet-Lab, allowing for switching between the currently used DUMMYNET for backwards compatibility and NETEM – in order to let experiments choose the best-suitable approach for their specific requirements.

REFERENCES

[1] International Telecommunication Union, "Open Systems Interconnection – Base Reference Model," ITU-T, Recommentation X.200, Aug. 1994.

[2] C. Raiciu, M. Handley, and D. Wischik, "Practical Congestion Control for Multipath Transport Protocols," University College London, London/United Kingdom, Tech. Rep., 2009.

[3] B. White, J. Lepreau, L. Stoller, R. Ricci, S. Guruprasad, M. Newbold, M. Hibler, C. Barb, and A. Joglekar, "An Integrated Experimental Environment for Distributed Systems and Networks," Boston, Massachusetts/U.S.A., Dec. 2002, pp. 255–270.

[4] A. Bavier, N. Feamster, M. Huang, L. Peterson, and J. Rexford, "In VINI Veritas: Realistic and Controlled Network Experimentation," in *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, Pisa/Italy, 2006, pp. 3–14.

[5] T. Friedman, S. Fdida, P. Duval, N. Lafon, and X. Cuvellie, "OneLab: Home," 2009.

[6] R. Steinmetz, J. Eberspächer, M. Zitterbart, P. Müller, H. Schotten, and P. Tran-Gia, "G-Lab Phase 1 - Studien und Experimentalplattform für das Internet der Zukunft," White paper, www.german-lab.de, Jan. 2009, available online (16 pages).

[7] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "PlanetLab: An Overlay Testbed for Broad-Coverage Services," *SIGCOMM Computer Communication Review*, vol. 33, no. 3, pp. 3–12, 2003.

[8] M. Carbone and L. Rizzo, "Dummynet Revisited," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 2, 2010.

[9] L. Rizzo, "Dummynet: A Simple Approach to the Evaluation of Network Protocols," *SIGCOMM Computer Communication Review*, vol. 27, no. 1, pp. 31–41, 1997.

[10] M. Hibler, R. Ricci, L. Stoller, J. Duerig, S. Guruprasad, T. Stack, K. Webb, and J. Lepreau, "Large-scale Virtualization in the Emulab Network Testbed," in *Proceedings of the USENIX Annual Technical Conference on Annual Technical Conference*, Boston, Massachusetts/U.S.A., 2008, pp. 113–128.

[11] S. Hemminger, "Network Emulation with Netem," in *Proceedings of the Linux Conference Australia (LCA)*, April 2005.

[12] B. Hubert, T. Graf, G. Maxwell, R. van Mook, M. van Oosterhout, P. B. Schroeder, J. Spaans, and P. Larroy, "Linux Advanced Routing and Traffic Control HOWTO," 2010.

[13] G. Kurtsou, "Layer 2 FreeBSD Dummynet Patch," 2009.

[14] B. Fink, "Manpage of NUTTCP," 2007.

[15] M. Carbone and L. Rizzo, "Adding Emulation to Planetlab Nodes," in *Proceedings of the ACM CoNEXT Workshop*, Rome/Italy, December 2009.

[16] J. J. Dongarra and T. Dunigan, "Message-Passing Performance of Various Computers," Knoxville, Tennessee/U.S.A., Tech. Rep., 1995.

[17] L. Nussbaum and O. Richard, "A Comparative Study of Network Link Emulators," in *Proceedings of the Spring Simulation Multiconference (SpringSim)*, San Diego, California/U.S.A., 2009, pp. 1–8.

 ISBN:978-1-61208-113-7

User utility function as Quality of Experience(QoE)

Manzoor Ahmed Khan
DAI-Labor Technical university Berlin, Germany
Email: manzoor-ahmed.khan@dai-labor.de

Umar Toseef
ComNets university of Bremen, Germany
Email: umr@comnets.uni.de

***Abstract*—The realization of a *user-centric* paradigm will revolutionize future wireless networks. For this innovative concept to materialize, a paradigm shift is required from a *long-term contractual* based service delivery to a *short-term contractual* and *dynamic service delivery* concept. However this necessitates that translation of user satisfaction into network technical indices, commonly termed as *Quality of Experience (QoE)*. In this paper we propose the *user utility function* to capture her satisfaction for different services. We validate the proposed utility function by extensively carrying out the simulations for VoIP, video streaming and FTP applications using OPNET simulator. We also suggest three different types of users on the basis of *user preference*.**

***Keywords*-utility function; user satisfaction; QoE;**

I. Introduction and Motivation

The business models of telecommunication operators have traditionally been based on the concept of the so-called closed garden: they operate strictly in closed infrastructures and base their revenue-generating models on their capacity to retain a set of customers and effectively establish technological and economical barriers to prevent or discourage users from being able to utilize services and resources offered by other operators. After the initial monopoly-like era, an increasing number of (real and virtual) network operators have been observed on the market in most countries. Users benefit from the resulting competition by having a much wider spectrum of choices for more competitive prices.

In its most generic sense, the user-centric view in telecommunications considers that the users are free from subscription to any one network operator and can instead dynamically choose the most suitable transport infrastructure from the available network providers for their terminal and application requirements [7]. One envisions that in future telecommunication paradigm, the decision of interface selection will totally be delegated to the mobile terminal enabling end users to exploit the best available characteristics of different network technologies and network providers, with the objective of increased satisfaction. The generic term satisfaction can be interpreted in different ways, where a natural interpretation would be obtaining a high quality of service (QoS) for the lowest price. In order to more accurately express the user experience in telecommunications, the term QoS has been extended to include more subjective and also application-specific measures beyond traditional technical parameters, giving rise to the quality of experience (QoE) concept. QoE reflects the collective effect of service performances that determines the degree of satisfaction of the end user, e.g., what the user really perceives in terms of usability, accessibility, retainability, and integrity of the service. Until now, seamless communications is mostly based on technical network QoS parameters, but a true end-user view of QoS is needed to link between QoS and QoE. While existing 3GPP or IETF specifications describe procedures for QoS negotiation, signaling, and resource reservation for multimedia applications, such as audio/video communication and multimedia messaging, support for more advanced services, involving interactive applications with diverse and interdependent media components, is not specifically addressed. Such innovative applications, likely to be offered by third-party application providers and not the operators, include collaborative virtual environments, smart home applications, and networked games. Perceived quality problems in future internet might lead to acceptance problems, especially if money is involved. For this reason, the subjective quality perceived by the user has to be linked to the objective, measurable quality, which is expressed in application and network performance parameters resulting in QoE. Technical Report 126 of the DSL Forum (Digital Subscriber Line Forum) is a good source of information on QoE for three basic services composing the so-called triple play services. One way to achieve QoE assessment is to perform subjective tests with panel of humans, which is not an attractive solution in online optimization, another approach may be to use objective testing to predict the MOS value of a service. However such approaches require original signals (for real-time applications e.g., ITU-T objective measurement standards like PESQ, E-model etc.) and are computationally complex. In addition these approaches do not capture user-satisfaction (user-preferences) based on the non-technical or economical parameters specifically pricing, reputation of operators etc.

Several research contributions on meeting the user QoS and bandwidth requirements are present in the literature, most of them mainly focus on homogeneous service demands, fairness etc. by suggesting *radio resource management* schemes and *scheduling algorithms* [17].[6] compares several scheduling algorithms for real-time and non-real time applications, similarly [14] suggests the QoS aware packet scheduling for real time multi-media traffic, and a simple priority order queue mechanism for non-real time applications. [9], [15] discuss the problem of utility based throughput allocation and load balancing, where the earlier reference restricts the utility to linear behavior, and the later formulates the objective as network wide utility function balancing network throughput and load distribution. User-centric network selection approaches based on various approaches including *policy based*, *fuzzy logic based* etc. are discussed in [2], [3], [10], [4]. However most of the research literature either formulate the network selection problem as a static optimization problem or theoretically assume that user satisfaction function for any application follows a function, and these assumptions are not supported by the validation that represents the realistic user satisfaction.

To address these issues we propose users utility function, that captures user satisfaction for *real-time* and *non-real-time* applications with respect to both *technical* and *non-technical*

 ISBN:978-1-61208-113-7

attributes. The estimated MOS outcome from utility function can then be used for network selection decision making. *The user network selection decision making can be based on maximizing the utility function and user utility also drives the operator strategies*. We in this work validate the proposed utility function by comparing MOS value curves attained from the proposed utility function to the ones obtained from objective measurement techniques and study the relationships between them.

II. Proposed utility function

We capture the user satisfaction using utility function, the term *utility* comes from the field of Economics. Utility is an abstract concept and is derived largely from Von Neumann and Morgenstern [11]. It is designed to measure the user satisfaction. A *utility function* measures users relative preference for different levels of decision metric attribute values. Thus preference relation can be defined by the function, say $U : X \longrightarrow \Re$, that represents the preference for all x and $y \in X$, if and only if $U(x) \geq U(y)$. Basically a utility function should satisfy non-station and risk aversion properties [11].

Let $U_i(b_{k,c}, Q_{k,c}, \pi_{k,c})$ represents the utility function of user i, then: $U_i(b_{k,c}, S_{k,c}, \pi_{k,c}) :=$

$$v_i(b_{k,c}) \prod_{l \in L} u_{il}(t_{c,k})^{w_l} . \pi_{c,k} + \sum_{j \in J} w_j u_{ij}(Q_{c,k}) \qquad (1)$$

User utility function is the function of offered bandwidth $b_{k,c}$, offered associated satisfaction attributes S_k^c (where $S_k^c = \{Q_{c,k}, t_{c,k}\}$), and the service price $\pi_{k,c}$. Here $k \in \Theta$ represents the finite set of user types (We consider three types of users namely Excellent, Good, and Fair). $w_j \in J$, $w_l \in L$ represents the weights of parameter j and l, these attributes are detailed in later section.

We decompose the user utility function into four components, namely i) bandwidth dependent utility component, ii) associated dependent attributes utility component, iii) associated independent attributes utility component, and iv) price dependent utility component.

A. Bandwidth dependent utility - Availability of bandwidth / transmission data rate plays a key role in evaluating the user QoE, therefore most of the literature work focuses on throughput optimization. However amount of bandwidth is strictly driven by the application types and user preferences. Application specific bandwidth requirements are well studied in the literature and standards documentation, however user preferences over the bandwidth requirements is a subjective quantity and depends on the type and context of users. In this connection, we characterize the proposed user types as; i) Excellent users - the users who prefer quality more than the service price, ii) Good users - the users who stand mid-way between the quality and price, and iii) Fair User - the users who values the service cost more than the service quality, indexed by k. The bandwidth dependent utility component explicitly captures user satisfaction for offered bandwidth values to different user and service types and is given by:

$$u_i(b_k^c) = \begin{cases} 0 & \text{if } b_k^c < \underline{b}_c^k \\ \mu_{k,c} \frac{1 - e^{-\beta_c (b_k^c - \underline{b}_k^c)}}{1 - e^{-\beta_c (\overline{b}_k^c - \underline{b}_k^c)}} & \text{if } \underline{b}_c^k < b_k^c < \overline{b}_k^c \\ \mu_{k,c} & \text{if } b_k^c \geq \overline{b}_k^c \end{cases} \qquad (2)$$

where $\mu_{k,c}$ is the maximum achievable MOS for the service class c, and user k. β_c represents the sensitivity of application c towards the amount of bandwidth, i.e. $\beta_{real-time-application} > \beta_{non-real-time-applications}$. The value of β is scaled between the value range $[0, 1]$, and for different k type users, $\overline{b}_{excellent} > \overline{b}_{good} > \overline{b}_{fair}$ and $\mu_{excellent} > \mu_{good} > \mu_{fair}$.

B. Associated dependent attributes utility - The term *dependent* here refers to the dependency on the bandwidth. This component of the user utility function is the function of *delay* and *packet loss* QoS metric parameters. Since both the mentioned parameters can be normalized into the *the lower the better* expectancy, therefore we capture the user satisfaction for these parameters as:

$$\prod_{l \in L} u_{il}(t_{c,k})^{w_l} := \left(\begin{cases} 1 & \text{if } l_{k,c} < \bar{l}_{k,c} \\ e^{l_{k,c} \zeta_{k,c}(l)} & \text{if } l_{k,c} \geq \bar{l}_{k,c} \end{cases} \right)^{w_{l,c}} \qquad (3)$$

where $t_{k,c}$ represents the finite set containing $l \in L$ dependent variables namely *delay* and *packet loss*. $\zeta_{k,c}(l)$ represents the sensitivity of user satisfaction towards the increasing values of l. $w_{l,c}$ (driven by the application type) represents the weighted contribution of utility degradation introduced by attribute l. $\bar{l}_{k,c}$ is the ideal attribute values for which the user k for any class of service c has the maximum achievable utility.

C. Associated Independent attributes utility - This component of user utility is the function of various attributes like *reputation of operator*, *security*, *battery life* etc. These attributes are of diverse scope and can be normalized on expectencies of *the lower the better*, *the higher the better*, or *the nominal the better*. User satisfaction for this component is purely attribute dependent i.e., the decision of using linear, exponential, lograthmic functions and control parameters depend on the attribute under consideration e.g., for security parameter, a function like bandwidth dependent utility may be used.

D. Price based utility - In addition to technical, user satisfaction is also influenced by the economical parameters. One can not neglect the importance of this parameter in decision making for network selection, when it comes to cost-sensitive user types (e.g., fair users). We capture the satisfaction of different user types with respect to service prices as: $u_k(\pi_{k,c}) = \tilde{\mu}_{k,c} - \frac{\tilde{\mu}_{k,c}}{1 - e^{\tilde{\pi}_{c,k}}} e^{-\pi\epsilon}$, where $\tilde{\mu}_{k,c}$ represents the maximum satisfaction level of user type k, and $\tilde{\pi}^{c,k}$ is the private valuation of service by user, and ϵ represents the price sensitivity of user.

III. Experimentation and utility validation for different services

A. Real-time VoIP applications

Streaming and conversational traffic classes can be combined in real-time applications, which are commonly termed as *inelastic* or *rigid* applications. Generally real-time applications are constrained by minimum amount of bandwidth i.e., application is admitted only when the demand for minimum required bandwidth is met. Such stringent requirement on bandwidth are represented by step like function, which results in a very narrow transition region between the two states (fully satisfied, unsatisfied). Such transition region is captured by the value of β of user utility function given in equation-2. This transition region represents

 ISBN:978-1-61208-113-7

very narrow required bandwidth range for different real-time applications e.g., audio broadcasting demands $60-80Kbs$, video broadcasting demands $1.2Mbs-1.5Mbs$ with MPEG1 coding standard.

A.1. VoIP objective measurement - In order to capture user satisfaction using simulation measurement methodology, we set up a simulation scenario with heterogeneous wireless technologies, and run a lengthy rounds of simulations to analyze the user satisfaction for different values of delay and packet losses, when she is associated to different codecs using ITU-T PESQ and modified E-models standard models. It should be noted service affecting factors [1] in addition to delay and packet loss are out of the scope of the objective measurements.

A.1.1. OPNET simulation setup - The components involved in the simulation setup include; i) impairment entity - we develop an impairment entity that introduces specified packet delay, packet loss and is also able to limit bandwidth available to a voice communication by performing bandwidth shaping using token bucket algorithm. ii) LTE radio access network, iii) WLAN radio access network, and iv) transport network. The simulation is setup such that the impairment entity resides between the caller and the callee, and introduces various delays and packet losses during the life of a VoIP call.

Note - The packet delay values in the simulation include only codec delay and transport network delay excluding fixed delay components e.g., equipment related delays, compression decompression delays and other internetwork codec related delays etc.

A.1.2. Simulation Results - We analyze the results for three different codecs namely i) G.711, ii) GSM EFR, and iii) G.729, which are characterized by their data rates. Each codec in a lossless (lossless is an ideal scenaio, where packetloss and delay values are idealy zero.) condition achieves the maximum MOS, $\overline{MOS}_c$, such that $\overline{MOS}_c \neq \overline{MOS}_{\tilde{c}}$. This characteristic of codec dictates that a user, when associated with a codec c, will have lossless MOS equal to $\overline{MOS}_c$ unless he is switchedover to the codec $\tilde{c}$. Codec switchover results in step-function like $\overline{MOS}$ value of user in a lossless scenario. This is depicted in Figure-1, which clearly shows that even in a lossless scenario, codec switchover introduces a marginal gain or loss in the MOS value. We now discuss a more realistic scenario, where user satisfaction is influenced by the packet loss and delay values, i.e. user associated with a specific codec c, with the MOS, $\overline{MOS}_c$, experiences delay and packet loss in the communication system. The consequence of system impairments is a degraded service, which in turn has negative impact on user satisfaction. In simulations, we use impairment entity to introduce customized delays and packet losses in the system and study the impact of parameter values on user satisfaction. Figure-2 shows the impact of delay and packet loss values on user satisfaction for $G.711$, $G.729$, and GSM EFR codec, As can be seen that all the codecs lead to different MOS values for different values of packet loss and delays.

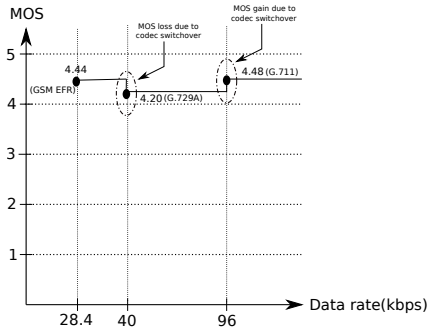


Figure 1. VoIP MOS for Loss-less Scenario with different codecs

Table I
UTILITY CONTROL PARAMETER VALUES FOR VOIP AND FTP APPLICATIONS

Application	Codec/size	$v_i(b_{k,c})$	$\zeta(pl)$	$\zeta(d)$	w_{pl}	w_d
Voice	G.711	4.48	0.03	0.0075	0.75	0.25
	G.729	4.20	0.03	0.0075	0.75	0.25
	GSM EFR	4.44	0.075	0.0033	0.4	0.6
FTP	20Mb	5.0	0.99	0.0429	0.5	0.5
Video	JM	3.98	0.031	0.011	0.7	0.3

A.2. Proposed utility function for VoIP applications - Although the proposed utility function in equation-1 captures user satisfaction for both technical and non-technical aspects, we limit here the scope of utility function to the technical part only so that we can validate it against the results obtained from the objective measurements. Since the objective measurements are carried out for different codecs and different packet loss, and delay values, therefore first two components of the proposed user utility (equation-1) are adequate to capture user satisfaction.

$$U_i(b_{k,c}, S_{k,c}) := v_i(b_{k,c}) \prod_{l \in L} u_{il}(t_{c,k})^{w_l} \quad (4)$$

where $v_i(b_{k,c})$ represent codec data rate, and this utility is given by a step like function. $v_i(b_{k,c})$ is tuned by the $\prod_{l \in L} u_{il}(t_{c,k})^{w_l}$ utility component. The control parameters for this utility component take different values for different codecs, which are given in the Table-I.

A.3. Validation - In this section we validate the proposed utility function for VoIP application by comparing the plots attained from the utility function to the plots we get from objective measurements (simulation results), this is shown in Figure-3. As evident from the figure that most of the points overlap well i.e., few points map exactly, for few MOS values the proposed utility function partially underestimates or overestimates the objective MOS values. This is further elaborated in Figure-4, where the correlation clearly strengthens the claim that proposed utility function for VoIP applications estimates the user satisfaction with appreciable confidence level.

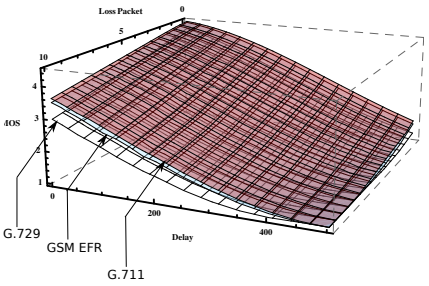


Figure 2. MOS values for different codecs

B. Non-real-time applications

Interactive and Background traffic classes can be combined in *non-real-time* applications, which are commonly termed as *elastic* applications, these applications are further divided into *symmetric* and *asymmetric non-real-time* applications. Generally *non-real-time* applications do not have stringent requirements

 ISBN:978-1-61208-113-7

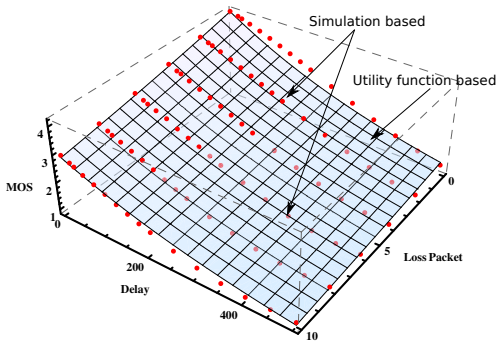


Figure 3. Mapping of objective and utility function measurements for VoIP application

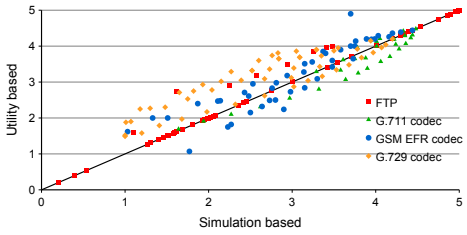


Figure 4. Correlation of objective and utility function for VoIP application

for bandwidth and delay. Such application can run even with minimal amount of available bandwidth, therefore, call admission control may not be needed in this case. TCP based non-real time applications ensure the error free delivery on the cost of waiting time introduced by TCP ARQ mechanism. This necessitates the proper investigation of *how much is the influence of packet loss and packet delay on achievable TCP throughput?* According to literature, TCP throughput is inversely proportional to round trip time of a network. In case of negligible packet loss rate following relation holds i.e., throughput $\leq$ (TCP buffer size) / RTT, where RTT is TCP segment round trip time. But if there are considerable packet losses than following relation holds, i.e. throughput $<$ (MSS/RTT)*(1/sqrt(PLR)), where MSS is the maximum TCP segment size, RTT is the round trip time and PLR is the packet loss rate. However above relations only show the upper bound of achievable TCP throughput. In order to investigate the more concrete throughput values of TCP in the presence of certain packet delay and packet loss rate, extensive simulations runs are provisioned.

User satisfaction metric for Non-real-time application - The performance metric to measure user satisfaction for non-real-time applications include *throughput*, *download response time* [13], and *MOS*. Although different, these performance measuring parameters are correlated. In this paper, we choose the MOS value as a performance metric for FTP applications. The motivation for selecting MOS as the performance metric is to have a generalized and common metric for different services.

B.1. FTP objective measurements - On the similar lines to the VoIP application objective measurmenets, we set up simulation scenario with heterogeneous wireless technologies and run lengthy rounds of simulations to analyze the user satisfaction for different values of delay and packet loss.

B.1.1. Simulation Settings and Methodology - This simulation environment also involves the *impairment entity*, and *LTE* in the similar fashion as discussed in the VoIP simulation settings, however in this case, the caller and the callee are replaced by the FTP server and FTP client. FTP server and client are connected through LTE access network. In our settings, an FTP client downloads a heavy file (of 20MB) through LTE access network. The choice of file size here is dictated by the facts; i) slow start effect of TCP can be ignored, ii) correlation of TCP throughput and distribution of packet losses within a TCP can be reduced. We artificially inject the packet delays by using the impairment entity, packet delays follow *Normal distribution*. A *bandwidth shaping* of 8Mbps is performed at a router in LTE transport network. We use the most widely used TCP flavor *New Reno* with receiver buffer size of 64KB. Moreover *windows scaling* option of TCP is disabled, *window scaling* option allows TCP maximum congestion window size to grow beyond 64KB. Due to deployment of accumulated acknowledgements, TCP is not very sensitive to the loss of few percent of acknowledgement packets in uplink direction, therefore, effect of packet loss is investigated only in downlink direction. It should also be noticed that processing delay and packet losses in network components (other than impairment entity) are negligibly small. Packet losses are injected based on Bernoulli distribution, packet delays are actually RTT values.

B.1.2 Simulation results - We analyze the impact of packet loss and delay values on the *user throughput* as shown in Figure-5. However *user throughput* does not directly show the user satisfaction, in this connection, we need to translate the *user throughput* values into *user satisfaction*. We carried such translation using the *throughput to MOS mapping approach* detailed below.

Throughput to MOS mapping - For such mapping we assume that a user of type k is subscribed to an amount of bandwidth b_k, such that the user remains fully satisfied (has the $MOS = \overline{MOS}$) as long as he receives the bandwidth b_k or $b_k + \epsilon$, and for any bandwidth less than b_k, the user satisfaction degrades and user reaches the *irritated state*, when the received bandwidth is $\underline{b}_k$. We term the bandwidth range $[\overline{b}_k - \underline{b}_k]$ as *feasible bandwidth range* for user k. This further necessitates a function of degradation and scaling the user satisfaction. We scale the user satisfaction for FTP applications on the same lines as in the case of VoIP MOS values i.e., [1-5], whereas the bandwidth dependent component of utility (equation-1) is used as the degradation function between fully satisfied and fully irritated states of user. For mapping the throughput results shown in Figure-5 into MOS scaled results, we set the control parameters of equation-2 to the following values; $\overline{b}_k = 1.265 \times 10^7 kbps$, $\underline{b}_k = 20250.449 kbps$, $\beta = 0.000006$, and $\alpha = 5$. The consequence of such mapping is depicted in Figure-6, which represents the *user satisfaction* in terms of MOS values for different achieveable data-rate values.

B.2. Utility function representation of FTP user satisfaction - We capture the user satisfaction for FTP application using the proposed utility function given in equation-1. From the simulations, what we get is the user throughput and impact of different packet loss and delay on the throughput as shown in Figure-5. The $v_i(b_{k,c}) \prod_{l \in L} u_{il}(t_{c,k})^{w_l}$ utility components capture the user satisfaction that is comparable to measurement results obtained from the objective testings.

 ISBN:978-1-61208-113-7

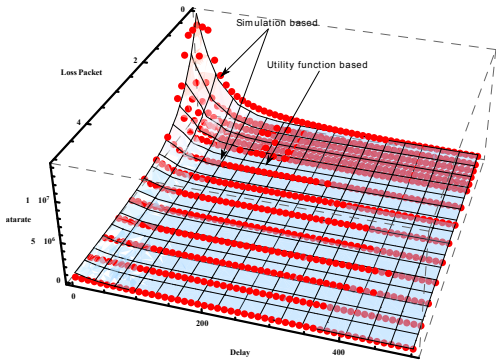


Figure 5. Overlapping of utility-based data rate values over the simulation-based data rate for FTP application

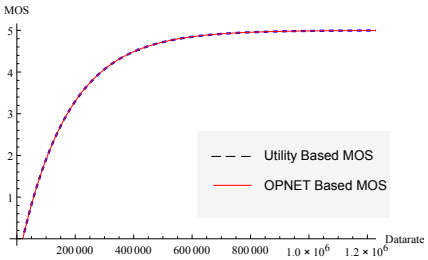


Figure 6. Comparison of Utility Based and Objective Measurement Based MOS values for FTP application

Let us first discuss the case, when we are not mapping the throughput over the MOS values, in this case $v_i(b_{k,c})$ component of the utility function takes constant value that represents the *throughput in lossless conditions* i.e., $\bar{b}_k = 1.265 \times 10^7 kbps$, which is shaped by the $\prod_{l \in L} u_{il}(t_{c,k})^{w_l}$ component of the utility function. The results attained from the utility-based measurements are overlapped over the objective measurement results, and it is observed that these map very well, as can be seen from the Figure-5. We also scaled the throughput by mapping it over the MOS (remember that similar parameter values for mapping in utility function based measurement i.e., $\bar{b}_k = 1.265 \times 10^7 kbps$,$\underline{b}_k = 20250.449 kbps$, $\beta = 0.000006$, and $\alpha = 5$ are used), the scaled result is presented in Figure-6. The results show that the utility function estimates the user satisfaction similar to the objective measurement and hence validate the proposed utility function.

C. Video streaming applications

In video streaming the most commonly used objective evaluations produce PSNR (peak signal to noise ratio) and Ssim (Structural similarity) as output video quality metrics.

PSNR - *PSNR defines the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. When comparing two video files, the signal is the original file and noise is the error which occurs due to compression or during transmission over the network. In the context of video quality evaluation, PSNR is taken as an approximation to human eye perception of image quality. It is measured in decibel units (dB).*

Ssim *- The Structural Similarity (Ssim) index is a novel method for measuring the similarity between two images. It takes the original undistorted image as a reference and provides the quality measure of the compressed/distorted image. Ssim index value ranges from -1 to 1. The higher the Ssim index value the higher the similarity between the two comparing images. For videos Ssim index is computed image by image [16].*

C.1. Methodology and simulation setup - In this work, we use PSNR as video quality metric owing to its widespread use in scientific literature. We calculate MOS value based on PSNR value. There are several parameters which decide the sensitivity of end user video quality to network impairments, such as: i) Type encoding - It is due to the fact that encoding schemes differentiate among frames based on their importance in the decoding process. Hence a loss of more important key frames deteriorates reconstructed video quality much more than the loss of less important non-key frame. ii) Error concealment method, iii) Frames per second, iv) MTU size of transport network, and v) Pre-filtration of codec etc. However a thorough study of the impact of all other above parameters on video file transmitted over a wireless network in the presence of additional IP impairments is beyond the scope of this work. We consider a reference video sequence called *Highway* for this work. The motivation to use this video sequence its repeated reference in a large number studies in video encoding and quality evaluation e.g. *Video Quality Experts Group*[5]. This video sequence has been encoded in $H.264$ format using the JM codec [14] with CIF resolution (352×288) using a target bit rate of $256 kbps$. $H.264$ codec has been selected because its widespread use can be seen in future communication devices. The reference video sequence has total 2000 frames and frame rate of $30 fps$. Key frame is inserted after every $10th$ frame which provides good error recovery capabilities. An excellent video quality is indicated by $38.9 dB$ as an average PSNR value of encoded video sequence. The video file is transmitted over the IP network considering MTU size of 1024 bytes. At the receiving end, video file is reconstructed from received IP packets. The reconstructed video file might have errors due to packet losses and delays in the transport IP network. Results presented in this work have been taken from OPNET simulation setup.

C.1.1 OPNET simulation setup - This simulation setup has two parts, the first part includes implementation of E-UTRAN, EPC network entities of LTE access network, and the second part of the simulation set-up is derived from EvalVid [8]. EvalVid is a framework which can be used for video quality evaluation. It provides both PSNR as well as MOS values of reconstructed video file. The motivation to use Evalvid is its flexibility to be used in conjunction to simulation environments like ns-2 and OPNET. None of the other available video evaluation tools provide such an interface. Target of this task is to get video quality metric for video file which is transmitted over LTE access network. The transport network part of LTE artificially introduces IP impairments to the transmitted video file. Here the IP impairment entity uses *Normal distribution* for packet delays and packet delay variations. This choice is based on empirical study of big IP networks. Moreover packet losses are injected using *Bernoulli distribution*.

Following sequence of action leads to video quality metric for a particular value of mean packet delay and packet loss rate. EvalVid tools are used to generate a file which includes information about packets (e.g. *packet type*, *size*, *count* etc). These are the packet which would carry video frames if video file is transmitted over an IP network in real world scenario.

 ISBN:978-1-61208-113-7

Packet size and *type* information is used to transmit the same number, type and size of packets over LTE access network using OPNET simulator. IP impairment entity injects specified *packet delays* and *packet loss rate* in the above generated packet stream. Associated information of received packets (e.g. *packet end-to-end delay*, *jitter*, *type* and *sequence number of lost packets*) is used to reconstruct video file. This task is performed using EvalVid tools. *Play-out buffer length* of $250ms$ is used in this step. The reconstructed video file is then compared against the raw formatted reference video file to compute PSNR values frame by frame. It tells about the noise produced by both *encoding* as well as *transmission errors*. Video quality metric is computed by evaluating the difference between quality of $H.264$ encoded video file and reconstructed video file. In MOS of every single frame of the reconstructed video file is compared to the MOS of every single frame of the reference video file. In the end average MOS value of whole video file is output. For PSNR to MOS translation following table is used [12]. The simulation results are shown in Figure-7, depicting the *video user* satisfaction for different packet loss and delay values.

C.3. Utility function representation of video user satisfaction - We capture the user satisfaction for video streaming application using the proposed utility function given in equation-1 on very similar lines to that of VoIP applications(for details, refer to VoIP application section) The first two components of user utility function estimate the user satisfaction for different values of packet loss and delays. For video application, the control parameters of the proposed utility function are listed in Table-I. In order to validate the proposed utility function, we overlap the results obtained from the utility-based measurement over the simulation-based measurements. It is observed that the proposed utility function estimates the *video user* satisfaction very similar to the satisfaction values from experimentation, this is evident from the Figure-7.

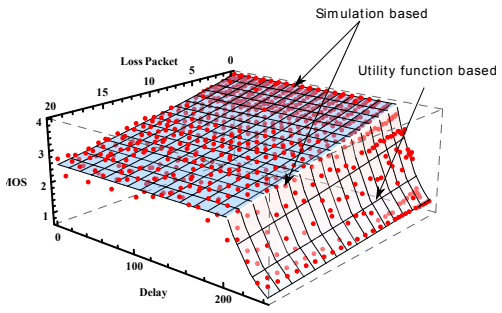


Figure 7. Overlapping of utility-based MOS values over the simulation-based MOS values for video application

Given the validation results in Figures-4,6,5,7, we confidently claim that the proposed utility function estimates the user satisfaction similar to subjective and objective measurements.

IV. Conclusion

In this paper we translated the user satisfaction in utility function. The proposed utility function captures user preferences over both technical and non-technical decision attributes. We carried out the extensive simulations to compute the user satisfaction for different service types including voice, FTP, and video streaming. Factors effecting the user perceived service quality have been discussed in detail for different test scenarios. We proposed the utility function that estimates the user satisfaction for different applications, and also validated the proposed utility function by comparing the utility-based results against the results attained from the objective measurements. We plan to extend this work to model operator utilities and find the local and global optimum solution for resource allocation at the operator level and network selection strategies (using proposed utility function) in different environments at user level.

References

[1] J. Fajardo, F. Liberal, I. Mkwawa, L. Sun, and H. Koumaras. Qoe-driven dynamic management proposals for 3g voip services. *Computer Communication*, 33, September 2010.

[2] V. Gazis, N. Houssos, N. Alonistioti, and L. Merakos. On the complexity of always best connected in 4g mobile networks, 2003.

[3] X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agusti, and F. Casadevall. Radio resource management in heterogeneous networks. In *Proceedings of the 3rd International Working erogeneous Networks*, 2005.

[4] L. Giupponi, R. Agusti, J. Perez-Romero, and O. Sallent. A novel joint radio resource management approach with reinforcement learning mechanisms. In *24th IEEE Ineternation Conference on Performance, Computing, and Communications*, 2005.

[5] Video Quality Experts Group. http://vqeg.org (last accessed september 2, 2010).

[6] P. Jos and A. Gutirrez. Packet scheduling and quality of service in hsdpa, October 2003.

[7] T. G. Kanter. Going wireless, enabling an adaptive and extensible environment. In *Mobile Network Applications*, 2003.

[8] J. Klaue, B. Rathke, and A. Wolisz. Evalvid - a framework for video transmission and quality evaluation. In *In Proc. of the 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, pages 255–272, 2003.

[9] X. Liu, E.K.P. Chong, and N.B. Shroff. A framework for opportunistic scheduling in wireless networks. *Computer. Networks*, 41.

[10] K. Murray and D. Pesch. Policy based access management and handover control in heterogeneous wireless networks. In *60th Vehicular Technology Conference*, 2004.

[11] John Von Neumann. Theory of games and economics behavior.

[12] Jens rainer ohm. bildsignalverarbeitung fuer multimedia-systeme, skript.

[13] ITU-T recommendations G.1030. Estimating end-to-end performance in ip networks for data applications. In *Series G: Transmission system and media digital system and netowrks*.

[14] S. Shin, S. Bahng, I. Koo, and K. Kim. Qos-oriented packet scheduling schemes for multimedia traffics in ofdma systems. *4th International Conference on Networking*, 2005.

[15] H. Wang, L. Ding, P. Wu, Z. Pan, N. Liu, and X. You. Dynamic load balancing and throughput optimization in 3gpp lte networks. In *IWCMC*, 2010.

[16] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. In *In proceedings on IEEE Transaction Image Processing,*, 2004.

[17] D. Zhao, X. Shen, and J.W. Mark. Radio resource management for cellular cdma systems supporting heterogeneous services. *IEEE Transactions on Mobile Computing*.

 ISBN:978-1-61208-113-7

Tuning Self-Similar Traffic to Improve Loss Performance in Small Buffer Routers

Yongfei Zang[1], Jinyao Yan[2]
[1]Information Engineering School, Communication University of China, 100024 Beijing, China
[2]Computer Engineering and Networks Lab, ETH Zurich, CH-8092 Zurich, Switzerland
Emails: {zangyongfei123@163.com, jyan@ee.ethz.ch}

Abstract—**The issue of router buffer sizing is an important research problem and is still open though researchers have debated this for many years. The research method can be classified into two kinds: one is based on queuing theory, the other uses TCP as model. From the point of TCP model, many researchers concluded that buffer size can be significantly reduced. It's desirable that the buffers are so small that fast memory technology and all-optical buffering can be used. But queuing model with self-similar incoming traffic suggested that extremely large buffers are needed to achieve acceptable packet loss rate. In this paper, we will first exam the performance of non-TCP and self-similar traffic with small router buffers, and then address the question how to improve the packet loss rate performance for self-similar traffic. Through a combination of simulation and analysis, we found that packet arrivals' burstiness has a significant influence on loss rate performance. We further point out a simple and effective approach, which smoothes the packet injections to the network, to improve the performance of small buffers at Internet core router for self-similar traffic.**

Keywords-buffer size; TCP; self-similarity; traffic smoothing.

I. INTRODUCTION AND MOTIVATION

All Internet routers need buffers to hold packets when TCP connections back off due to the congestion of the network and buffer the transient bursts that naturally occurred due to the characteristics of strong bursts and self-similarity of the Internet traffic, so the router buffers can keep high utilization of output link and reduce the packet loss rate. Meanwhile, buffers introduce queuing delay and jitter, and increase the router cost and power dissipation inevitably. The issue of sizing router buffer properly has generated much debate in the past few years. Different assumptions and objects have led to different conclusion. However, some of the recent reasearch all claimed that the router buffer can be significantly reduced in some cases. It is a good news for manufacturers and the development of all optical routers considering that recent advanced technology can at best hold a few dozen packets in an integrated opto-electronic chip [5].

The rule-of-thumb commonly used by router manufacturers today was proposed by Villamizar and Song in 1994 [1]. It claims that in order to make full utilization of the bottle link, a router needs a bandwidth-delay production buffering because of the sawtooth-like of TCP's congestion control algorithm, i.e., B = RTT $\times$C, where C is the capacity of the bottleneck link, B is the buffer of the bottleneck router and RTT is the average round trip time of a single and persistent TCP flow that attempts to saturate the link. The amount of buffers is in direct proportion to C and it will be a very large value considering that nowadays that backbone links commonly operate at Gbps magnitude.

In 2004, Appenzeller et al. from Stanford University challenged role-of-thumb. They concluded that when a large number of long TCP flows go through a bottleneck link in the core of the network, the buffer requirement decreases with the square root of the number of long TCP flows [2], i.e., $B = RTT \times C/\sqrt{N}$. According to their conclusion, a core router carrying 10000 long-lived flows needs only 1% of buffers proposed by rule-of-thumb.

In 2005, Enachescu et al. showed that if the TCP sources are paced or the access network is much slower than the backbone, and the maximum widow size has upper bound, O(logW) buffers (a few dozen packets) are sufficient if we are willing to sacrifice a small amount of the link capacity (say 10-20%), where W is the window size of each flow [3]. This result has made a useful exploration for the building of all optical routers with small integrated optical buffers.

In 2007, authors of [4] used a different metric and parameter to revisit the issue of router buffer sizing. Instead of only focusing on aggregate metrics such as link utilization and packet loss rate, they used average per-flow throughput to assess TCP performance. They claimed that the ratio of output/input capacity at a network link largely determines the required buffers. If the ratio is larger than 1, the loss rate drops exponentially with the buffer size and the optimal buffer size is extremely small (a few packets in practice). Otherwise, if the ratio is lower than one, the loss rate follows a power-law reduction with the buffer size and significantly large buffering is needed, especially with long-lived TCP flows which spend most of their time in congestion-avoidance.

The sizing router buffer formulas above is concluded based on closed-loop TCP congestion control model. Statistics shows that about 90% Internet traffic are TCP based while the rest traffic is transmitted over UDP and considered as open-loop traffic. Authors in [15] examined the dynamics of UDP and TCP interaction at a core router with few tens of packets of buffering and discovered the anomaly of UDP traffic's loss performance. From the view of queuing theory with specific incoming traffic model, the buffer sizing for open-loop traffic is quite different from

 ISBN:978-1-61208-113-7

TCP. In the open-loop model, the router buffer is often modeled as a single queue with constant service rate (i.e., the capacity of the output link) and buffer size. The overflowing rate of the buffer depends on not only the buffer size and the capacity of the output link, but also the packet arrivals' patterns and the traffic's statistical features [6]. Various studies [7, 8, 9] have shown that network traffic exhibit ubiquitous properties of self-similarity. Analysis on video traffic, which is often transmitted over UDP, shows that self-similarity is also an inherent feature of VBR video traffic [10, 11]. The self-similar nature of network traffic has a significant influence on the queuing performance of router buffer. Authors of [12] pointed out that the packet loss rate in a network with self-similar traffic might be several orders of magnitude higher than that predicted by the traditionally used Markovian traffic models.

In this paper, we will exam how the burstiness of self-similar traffic affects the queuing performance in the condition of small router buffers, and propose methods to improve the performance of small buffers in Internet core router for self-similar traffic.

The rest of the paper is organized as follows. In Section II, we compare the loss performances of self-similar traffic with varied traffic burstiness with Poisson traffic. We study how the burstiness of data sources influences the queuing performance of router buffer. In Section III, real video traces from the Internet and CBR traffic are used to validate our finding. We summarize our work and point out directions for future work in Section IV.

II. THE BURSTINESS OF SELF-SIMILAR TRAFFIC AND PERFORMANCE

For self-similar traffic, bursts will exist across a range of scales and the positive correlations in traffic will adversely affect the QoS provided to network users [13]. Simply increasing the routers' buffer sizes will have marginal impact on the packet loss rate. The heavy tailed nature of the burst size distribution [11] implies that only extremely large buffers are effective in reducing packet loss rate [12]. The queuing delay introduced by large buffers will impact the transfer delay performance of delay-sensitive traffic such as streaming media.

To present how the traffic's self-similar influences router's queuing performance, we use NS2 simulator on the commonly used dumbbell topology to simulate self-similar traffic and Poisson traffic.

The aggregation of many On/Off sources with heavy-tailed ON periods exhibits Long-range Dependence (LRD) [14]. In our simulation, we aggregate many Pareto On/Off Traffic Generators in NS2 to generate self-similar traffic. We use Poisson traffic generator in NS2 to generate Poisson traffic. Because of Poisson process' additive property, we can use a single flow to represent the aggregation of many individual ones passing through the bottleneck link. UDP is used for both the self-similar and Poisson traffic. The capacity of the access links is 10Mbps, and the propagation delays on the access links uniformly distributed between [1, 25] ms. The capacity and propagation delay of the core link are 10Mbps and 50ms respectively. We employ FIFO queue with drop-tail queue management, which is commonly used in most router today. There are 100 On/Off source nodes each with the same configuration (burst_time_ 500ms, idle_time_ 500ms, rate_ 200Kbps, packetSize_ 200, shape_ 1.5). The mean rate during an ON-Off pair is 100Kbps. We set the Poisson rate to 10Mbps. So, in all simulations, the output link is lightly saturated. We examine the packet loss rates of self-similar traffic and Poisson traffic while increasing the buffer size at bottleneck link.

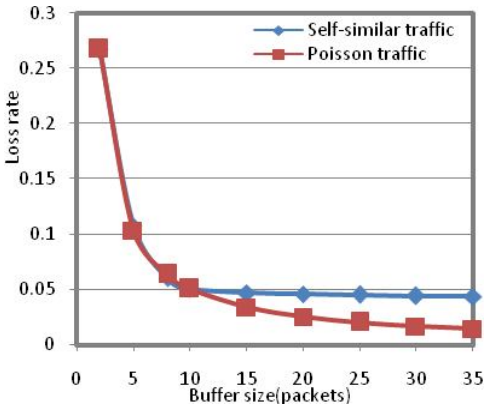


Figure 1. Loss rate for different buffer size

Figure 1 shows that in both cases, packets loss rate falls rapidly to a low value (5%, when the buffer size is 10 packets) as the buffer increases. After that, the self-similar traffic's loss rate curve drops very gently with the increase of the buffer size, while the loss rate of Poisson traffic falls faster than self-similar traffic. For self-similar traffic, increasing the buffer size simply will not get a good gain at loss rate.

One important reason for the self-similarity of network traffic is the statistical property of the size of the data blocks to be transferred, such as Web size, the size of Internet video's frames or GoPs. It's infeasible to change the statistical properties of the data to be transferred. From the observation on the loss rate performances' differences between the self-similar and the Poisson traffic, we consider the burstiness of packet arrivals leads to these differences. If we can make data sources to send data more smoothly, then what will happen?

In the next simulation, we will check the loss performance with different burstiness of self-similarity. We keep both the duration of On-Off pair (1000ms in our simulation) and the mean data rate at a constant value for self-similar traffic. By changing the length of On period ranging from 1ms to 1000ms, we adjust the burstiness for self-similar traffic. Let the mean data rate unchanged. Figure 2 shows the loss rate as a function of mean On time with buffer size 10 packets. We observe that the loss rate nearly falls exponentially with the increase of mean On time, which means that the data sources' burstiness has a significant influence on the loss performance. When the mean burst time is 1000ms, the Off time becomes to 0ms and each data source sends data with a low constant rate, namely with the lowest burstiness. Correspondingly, the loss rate achieves to the least value.

 ISBN:978-1-61208-113-7

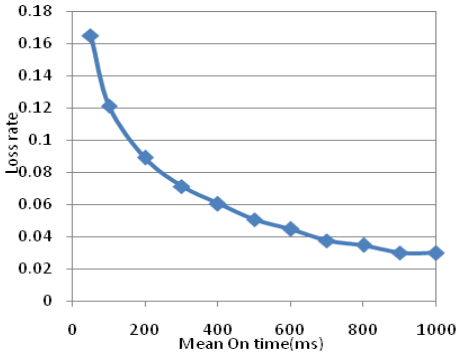


Figure 2. Loss rate for different mean On time

From the above, we can conclude that making the senders to space out packets evenly to weaken the burstiness of the data sources can improve the loss performance remarkably, even though the traffic to be transferred is self-similar. Therefore, we propose a simple and effective approach, which smoothes the packet injections to the network, to improve the performance of small buffers at Internet core router for self-similar traffic. For self-similar video streams, we send the frames with a constant rate continuously instead of sending the whole frame instantaneously as soon as we receive the frame from the application program. In the next section, we will validate our conclusion for co-existing TCP and UDP traffic with small buffer size routers.

III. SIMULATION VALIDATION

Let us consider a more realistic case of non-persistent TCP flows co-existed with UDP flows. We keep the fraction of UDP traffic fixed at about 8% in our simulations as that in the Internet. We use TCP traffic as background flows and focus on the loss performance of UDP traffic. Modified Harpoon system is used to generate closed-loop TCP flows. The size of TCP transfers follows Pareto distribution. After each download, an idle time which follows an exponential distribution with mean duration of 1 second follows until next TCP transfer starts [4]. Each TCP source can be seen as an On/Off model. The aggregation of many these TCP sources exhibit LRD property.

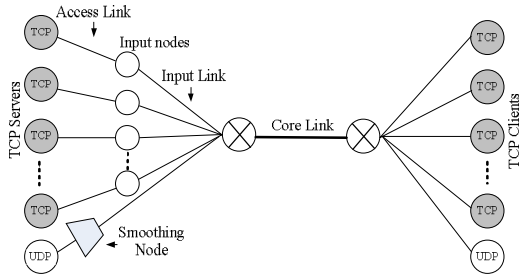


Figure 3. Simulation topology

Figure 3 shows our NS2 simulation topology. Our simulation setting has referred to [15], which focused on the anomalous loss performance of mixed real-time and TCP traffic. However, we simulate and compare the loss performance of the Internet video traffic and smoothed traffic over UDP. In [15], the TCP traffic was generated from persistent TCP flows while we use more realistic non-persistent TCP flows to generate it.

TABLE I. SIMULATION CONFIGURATION

Type \ Value \ Param		Buffer size (packets)	Link Capacity (Mbps)	Propogation Delay (ms)
Core Link		variable	10	50
TCP SACK	Access links	0	1	5,7,9…43
	Input links	100000	100	5
UDP	Input link	0	10, 0.256	10

Input link means the link that directly connects to the input core node.

TABLE I shows parts of simulation configuration. We employ FIFO queue with drop-tail queue management. There are 20 servers that are connected to 20 input nodes. Each of the 100 TCP users at client-side selects a server randomly to ecreate connections through the core link. The TCP transfer follows a Pareto distribution with mean 100KBytes and shape parameter 1.5. There are 3 UDP source nodes connecting to the input core node directly. The buffers of all the access links are too large to induce loss rate, so the output link is the single bottleneck. The simulation duration is 300s.The reported results ignore the first 20s of each simulation. TCP and UDP packet sizes were fixed at 1000Bytes and 200 Bytes respectively.

In what follows, we compare the loss performance of self-similar video traffic with that of smoothed video traffic. We insert a smoothing node between the UDP source nodes and the input core node with very large buffers (1 million packets) to ensure no dropped packet and 256 Kbps capacity of output link. The smoothing node buffers video trace packes and sends them smoothly to the input core node.

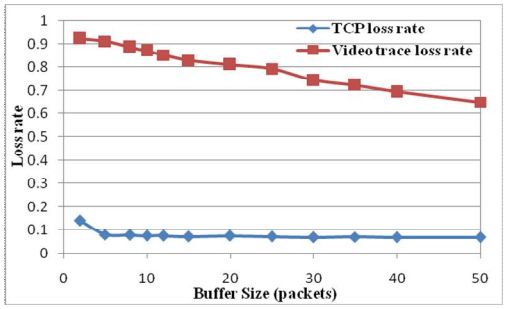


Figure 4. Video trace over UDP: Loss rate for different buffer size

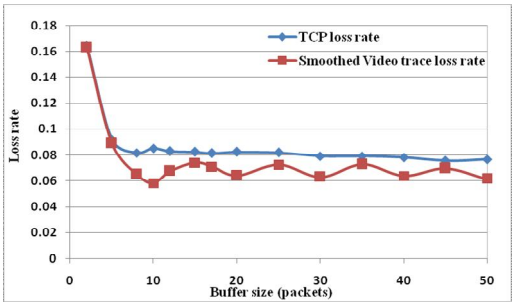


Figure 5. Smoothed Video trace over UDP: Loss rate for different buffer size

We use video traces of Jurassic Park I, Silence of The Lamb and Star Wars IV encoded by H.263 to generate UDP traffic [16]. Each video trace's mean rate is about 256 Kbps and the capacity of each UDP's input link is set to 10 Mbps. Previous study has shown that VBR video traffic is self-similar. So the UDP traffic generated from video traces is much bursty. We plot the bottleneck link utilization and loss rate of TCP and UDP as functions of the buffer size in Figure 4 and Figure 5.

We observe that loss rate of TCP are nearly identical. While there are vast differences between the UDP loss rate of Figure 4 and Figure 5. In Figure 4, the UDP loss rate drops linearly with the buffer size, while in Figure 5, the loss rate of UDP falls rapidly to a low value, then falls gently(but with high variance). For instance, the loss at 20 packets of buffering in Figure 4 is approximately 10 times higher than that of Figure 5.

We can explain with the findings in section II for the significant differences between the UDP loss rate curves in figures. In Figure 4, UDP traffic is generated from video traces. If the video frame to be transferred is larger than UDP packet size (200 Bytes in our simulation), the frame will be cut into a few of packets, then transferred simultaneously. But after smoothing, packets are sent in an approximately constant rate. So, UDP traffic generated from video traces is burstier than smoothed traffic though it originated from self-similar traffic. Therefore, our approach smoothing the packet injections to the network can improve the loss performance of small buffers at Internet core router for self-similar traffic.

IV. CONCLUSION AND FUTURE WORK

The study of sizing router buffer has generated much debate over the past few years. Researchers have questioned the commonly used rule-of-thumb which leads to a huge packet buffers in core routers today and have argued that small buffers at core routers are sufficient to meet acceptable performance. Various studies have shown that network traffic exhibit ubiquitous properties of self-similarity, from the point of queuing theory, extremely large buffer is needed. In this paper, we exploit how to improve the queuing performance of self-similar traffic at a bottleneck link router equipped with small buffers.

Through a combination of simulation and analysis, we found that there exists huge difference of loss performance between Poisson traffic and self-similar traffic due to the different bursty strength of packet arrivals. We can smooth packets injections at the edge of the network for self-similar traffic. Our realistic simulation mixed with TCP and UDP traffic shows that smoothed video traffic has a much better loss performance than VBR video streaming. We suggest that to adapt self-similar traffic to the small buffers at Internet core router, a simple and effective way to improve the queuing performance is smoothing the packet injections to the network.

As an important part of our future work, we will design the algorithm for our approach and implement the algorithm to test the performance in real network. After all, the "smoothing node" in our simulation was primary and used to make qualitative analysis. We will also revisit the issue of sizing router buffer with comprehensive consideration of TCP model and queuing model for self-similar traffic.

ACKNOWLEDGMENT

We would like to thank Yue Zhou and Botao Bai for their help in discussion and edit. Yongfei Zang is supported by NSFC under grant No. 60970127 and Key Project of Chinese Ministry of Education (No. 109029). Jinyao Yan is supported by Swiss National Science Foundation (No.200020_121753) and by Program for New Century Excellent Talents in Chinese University (NCET-09-0709).

REFERENCES

[1] C. Villamizar and C. Song, "High performance TCP in ANSNET," ACM Computer Communications Review, 24(5):45–60, 1994.

[2] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," In Proc. of the SIGCOMM 2004. New York: ACM Press, 2004. 281−292.

[3] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, "Routers With Very Small Buffers," Proc. IEEE INFOCOM, Barcelona, Spain, Apr 2006.

[4] R. S. Prasad, C. Dovrolis, and M. Thottan, "Router Buffer Sizing Revisited: The Role of the Output/Input Capacity Ratio," In ACM CoNEXT, USA, 2007.

[5] H. Park, E. F. Burmeister, S. Bjorlin, and J. E. Bowers, "40-Gb/s optical buffer design and simulation," Proc. Numerical Simulation of Optoelectronic Devices (NUSOD), California, USA, Aug 2004.

[6] I. Norros, "A Storage Model with Self-similar Input," Queueing System, vol. 16, pp. 387-396, 1994

[7] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," IEEE/ACM Trans. Networking, vol. 2, no. 1, pp. 1 - 15, 1994.

[8] V. Paxson, "Empirically-derived analytic models of wide-area TCP connections," IEEE/ACM Trans. Networking, vol. 2, pp. 316 - 336, 1994.

[9] M. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," Proceedings of the 1996 ACM SIGMETRICS Conference, Philadelphia, PA, pp. 160-169, May 1996.

[10] J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Long-range Dependence in Variable Bit-Rate Video Traffic," IEEE Transactions on Communications, Volume 43, pp. 1566-1579, 1995.

[11] M. Garrett and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," Proceedings of ACM SIGCOMM '94, London, UK, pp. 269-280, August 1994.

[12] Y. Chen, Z. Deng, and C. Williamson, "A Model for Self-Similar Ethernet LAN Traffic: Design, Implementation, and Performance Implications," Proceedings of the 1995 Summer Computer Simulation Conference (SCSC'95), Ottawa, Ontario, pp. 831-837, July 1995.

[13] N. Duffield, J. Lewis, N. O'Connell, R. Russell, and F. Toomey, "Predicting Quality of Service for Traffic with Long Range Dependence," Proceedings of ICC'95, Seattle, WA, pp. 473-477, September 1995.

[14] W. Willinger, M. Taqqu, R. Sherman, and D. V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," In ACM Sigcomm, 1995.

[15] A. Vishwanath and V. Sivaraman, "Routers With Very Small Buffers: Anomalous Loss Performance for Mixed Real-Time and TCP Traffic," Proc. IEEE IWQoS, The Netherlands, Jun. 2008.

[16] http://trace.eas.asu.edu/TRACE/ltvt.html, November 8, 2010

 ISBN:978-1-61208-113-7

Performance evaluation of Burst deflection in OBS networks using Multi-Topology routing

Stein Gjessing
Dept. of Informatics, University of Oslo & Simula Research Laboratory,
P-O. Box 1080, N-0316 Oslo, Norway. Email: steing@ifi.uio.no

Abstract – **This paper evaluates the combination of Optical Burst Switching (OBS) and Multi-Topology (MT) routing. Using MT routing, a source router has a choice of sending IP packets on several different paths to the destination. In OBS networks, deflection may reduce burst loss rate. We evaluate a deflection method that is based on MT routing and that ensures that deflected bursts will not loop indefinitely in the network. The performance of the method is evaluated by simulation and compared to two other deflection methods as well as just discarding burst that can not be scheduled. Performance is evaluated in three irregular networks with different topology characteristics. Our main load is IP-packets that arrive according to a self similar process. These packets are assembled into bursts that are transmitted either when the burst buffer is full or a timer expires.**

Keywords: Optical burst switching, Multi-Topology routing, Performance modeling, Burst loss rate, Burst deflection, Self-similar and Poisson arrival processes.

I. INTRODUCTION AND MOTIVATION

In Optical Burst Switched (OBS) networks [1], packets (e.g. IP packets) are assembled into bursts in the optical network ingress nodes, and the complete burst is transmitted either when the burst buffer is full or when a timer expires (hybrid burst assembly). A control packet precedes the burst in the network and reserves resources for the succeeding burst. In this paper "Just Enough Time" scheduling is used [2]. The burst is kept in the optical domain, while the control packet is converted from optical to electrical (and back) in each switch.

When the time slot that a burst needs on the output fiber is not completely available, there is a *contention* on the output line. The simplest approach is then to discard the burst. However, by deflecting the burst and send it out on another line, the burst may later arrive safely at its destination [4].

In general, deflection methods have three main drawbacks [5,6,7]: Some methods deflect bursts in a random direction, which might be counterproductive when considering the destination of the burst. The burst may also return to the point from which it was originally deflected, which may cause indefinite looping and even more contention. Some deflection methods try to deflect the packet on an alternative (loop free) path towards the egress. With connection-oriented routing, this will always be possible (given that the topology is bi-connected). In this paper we assume connection-less routing, and then an alternative loop free path may not be readily available [8].

Multi-Topology (MT) routing is developed within the Internet Engineering Task Force (IETF) [9]. MT routing is used in IP networks so that different streams, different traffic classes or different services (eg. multicast and unicast) can be forwarded in different topology images, and hence take different paths to the network egress node. These different topology images are subsets of the original topology. In each subset (topology) all routers are still present, but some links are removed. However, links should only be removed such that the network is still fully connected. The IETF "Request for Comments" for MT routing ([9]) specifies that a packet is forwarded in one and the same topology from ingress to egress. When used as a method for burst deflection, as will be described in the sequel, this will not be the case, but topology changes must be restricted in order to avoid indefinite looping.

In this paper we evaluate burst deflection in OBS networks based on MT routing by simulating traffic in three networks. The arrival process of bursts into the OBS-networks [10] is made up from self-similar IP-packets, and simulate a hybrid burs assembly method. At the end of the paper we compare these results to results achieved by an arrival process using Poisson distributed bursts.

The deflection method used in this paper was proposed in [20]. The contribution of this paper is a much more thorough discussion and evaluation of its performance.

This paper is organized as follows. In the next section we present MT routing and our deflection method based on MT routing that guaranties freedom from (indefinite) looping in any (bi-connected) topology. In section 3 we describe our performance evaluation method. In sections 4 we compare the performance of the different methods using three different irregular network topologies and self similar IP-traffic. In section 5 we compare our results with Poisson distributed burst arrivals. Finally in section 6 we conclude.

II. MULTI-TOPOLOGY BURST DEFLECTION

In an MT-capable IP-router there is (conceptually) one forwarding table for each topology image. One topology is the original topology, usually called the *default topology*, while in this paper we call the other topologies *backup topologies*. These backup topologies are subsets of the original topology, where some links are removed in each topology, while the network is still fully connected. In order to identify the topologies and the forwarding tables, the original (default) topology/table is numbered 0, and the backup topologies/tables are numbered from 1 and up.

MT routing is developed for shortest path, connection-less forwarding, and we assume that the switches in our OBS networks forward the bursts the same way. However, more sophisticated routing algorithms e.g. based on

 ISBN:978-1-61208-113-7

knowledge about the traffic matrix, may be used instead of shortest path routing (e.g. [11]).

All bursts are initially routed in the default topology. When the control packet that precedes the burst, arrives at a switch, the switch first tries to forward the burst (and the control packet) on the primary output link as decided by the default forwarding table. By installing wavelength converters, the probability of finding an available time slot increases [3]. In this paper we assume full wavelength conversion.

If there is a contention on the primary output link, the burst (and the control packet) is deflected according to one of the backup forwarding tables. As in MT routing, all switches contain one pre-calculated forwarding table for each topology. In order to be able to handle contention on any link, we need each link to be removed from at least one topology. We define a *complete* set of backup topologies as a set of topologies in which all links in the original topology are removed at least once. Figure 1 shows a full (the original network) topology on top left, and a complete set of 3 backup topologies.

We have devised algorithms to find complete sets of backup topologies for a given network [12,13]. The sizes of these sets have been shown to be surprisingly small; we have never come across a (normal) network that needs more than 5 backup topologies. When a complete set of backup topologies are found, each switch calculates one (loop free) forwarding table for each topology. In the network in figure 1 each switch will have four forwarding tables: One default forwarding table (according to the full topology) and three backup forwarding tables (These tables might be optimized to fill less space than four times what is needed for one table). MT routing does not specify what type of traffic maps to the different topologies, although the Type of Service (TOS) field in the IP header might be used if available.

In our burst deflection mechanism based on MT routing, a number in the control packet header tells which topology the burst (and the control packet) is currently forwarded in. Whenever a control packet arrives at a switch, this topology number is extracted and the corresponding forwarding table is used to find the bursts primary output link. If a switch can not send a burst out on a primary link because of contention, it deflects the burst by sending it out in any of the backup topologies that does not contain this link. Because of the way the complete set of backup topologies is constructed, at least one such topology does exist. The number in the control packet header is then set to this new topology number, and the burst is forwarded all the way to the egress in this topology (assuming no more deflections). If the burst experience a second (or third etc.) contention, it may conditionally be deflected once more. However, in order to avoid indefinite looping, we restrict all bursts to be routed in each topology at most once. This is achieved by only allowing a burst to be deflected to a topology with a higher number. When there is no higher numbered topology available, the control packet and the burst is discarded.

All backup topologies are fully connected, and loop-less forwarding tables are precomputed for all topologies. When deflected to another topology, the burst may return back to a node in the network it has visited before. However, because of the restriction that the burst is routed in each topology at most once, the burst will never loop indefinitely in the network.

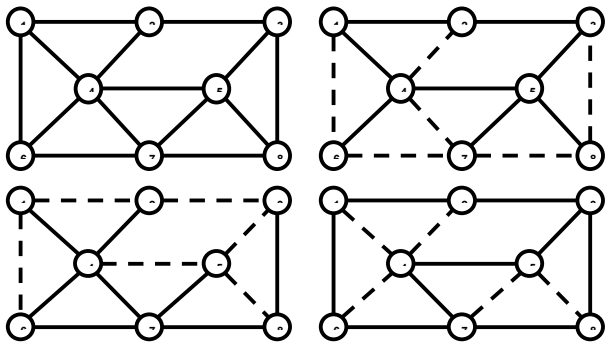


Figure 1. Original network top left, and a complete set of three backup topologies. Removed links are dashed. Notice that all links are dashed at least once.

III. PERFORMANCE EVALUATION

We have implemented a full OBS discrete event simulation model in the J-sim framework [14]. The data sources and burst assembly modules, as well as the OBS-switches and schedulers are built from scratch. Topologies, link propagation times and forwarding tables for the specific scenarios are read from files at system start up time.

The traffic load onto an OBS core network may come from IP-subnets and Ethernets. It is well known that Ethernet and IP traffic exhibit self similar properties [16]. We generate a self similar arrival process using a large number (100) of Pareto sources, with Hurst parameter 0.9, in each ingress node [17]. Whenever a Pareto source starts a new on period, a destination address is chosen according to the probability given by the traffic matrix, and IP packets are generated and sent to the same destination with constant intervals of 10 μs for the duration of the on period.

The size of the IP packets is varied from 80 to 1600 bytes, with a mean of 500 bytes. IP packets are assembled into bursts by a hybrid burst assembly method, meaning that a burst is transmitted when it is full, or a timer expires, whatever comes first (a 2 ms timer value is used in this paper). In this paper fixed burst size (50 000 bytes) is used.

When, at the end of the paper, we generate network load using Poisson distributed bursts, each ingress node runs one Poisson process per optical egress node, generating fixed sized bursts (50 000 bytes). The mean arrival rate is determined by load in the traffic matrix.

While the bursts are kept in the optical domain, and use very short time through a node, the control packet delay used in this paper is 10 μs in each node. The control packet lead time (CPT, i.e. how long ahead of the burst the ingress sends the control packet) is varied from 90 to 200 μs, depending on the diameter of the network (in number of switches). Hence, if a burst loops in the network, it will overtake the control packet (and they both become discarded) in between 9 to 20 hops. All experiments reported in this article are set up with equal capacity links.

 ISBN:978-1-61208-113-7

Each link has 10 channels (lambdas) and each channel has a capacity of 1 Gbit/sec.

We compare deflection based on MT routing (denoted ***Multi-Topology*** in the plots) with two other well known deflection methods: ***Hot Potato*** that chooses an alternative output link at random and ***Second Shortest*** path that tries to output the bursts to the output link where the next switch has the shortest distance to the destination (excluding the primary output link). Notice that for both methods a packet may be deflected back to where it came from, and hence in general these methods can not guarantee freedom from looping. Indefinite looping in the network is only prevented by the fact that when the burst is overtaking the control packet they are both discarded. In the case that the control packet is not able to reserve the needed resources for the data burst at all (deflection is not possible), the burst (and the control packet) is discarded by the switch.

In addition to comparing MT deflection with Hot Potato and Second Shortest, we also compare it with ***Regular*** burst dropping, i.e. when a burst may not be scheduled on the primary output link, it is immediately discarded (no deflection). The performance evaluation is carried out using three realistic and irregular networks with different characteristics; the Pan-European COST 239 network [18] and two networks from the Rocketfuel project from Washington University [15]; the Exodus network and the Sprint US network.

The COST 239 network is a proposed Pan-European core network topology consisting of 11 nodes (European cities) connected by 26 (bidirectional) links. The propagation delays are estimated based on the distances between the cities. The control packet lead time used by the ingress nodes is 90μs. The Exodus network is described by the Rocketfuel project and is AS number 3896. By collapsing switches in the same cities, and also collapsing parallel links, we have reduced the network to 17 nodes connected by 29 links. The link latencies vary from 2 to 15 ms. Initial control packet lead time is set to 120 μs. The second network from the Rocketfuel project is the Sprint US network (AS 1239). Also this network we have reduced, this time to 45 switches and 95 links. Link latencies vary from 2 to 64 ms. Initial CPT is set to 200 μs. All nodes are ingress nodes (generating traffic), egress nodes and internal switching nodes in the network. The traffic matrix is symmetric all-to-all. For each experiment we have made a one-second run for each load value.

IV. SIMULATION OF IP-TRAFFIC

In this section we report simulation results from running the three deflection methods, Multi-Topology, Hot Potato and Next Shortest as well as no deflection (Regular). The burst arrival process is a hybrid burst assembly of simulated self similar IP traffic.

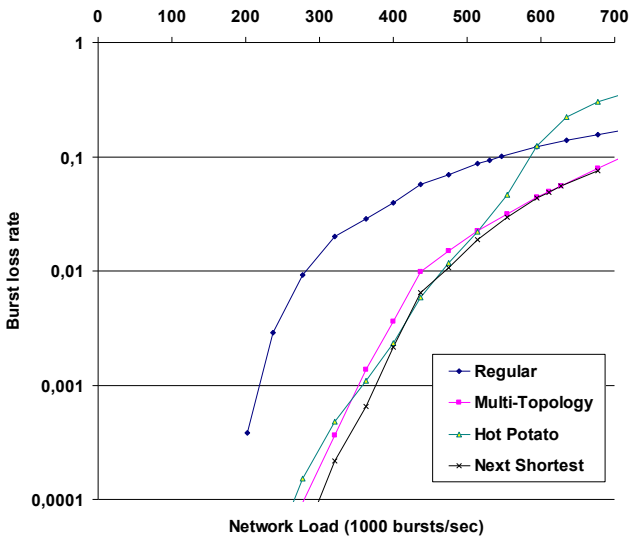


Figure 2. Burst loss rate in the Cost network with increasing network load.

A. *The Cost network*

The results are depicted in figure 2. With 100 Pareto sources per ingress node, the maximum load it is possible to send into the network from the 11 ingress nodes is approximately one M bursts/sec. As long as the total traffic generated is below 150 000 bursts/sec (i.e. each of the 11 ingress nodes generates about 5.5 Gbit/sec), there is no packet loss anywhere in the network.

When the load has increased to 300 000 bursts/sec, the burst loss rate for the Regular method is above 1%, while all the other methods still yield very good results. There is a very distinct change in the increase for the Hot Potato deflection method compared to the other methods at about 3% loss rate. Here the loss rate of this deflection method starts to increase steeply, while the loss rate of Multi-Topology and Next Shortest continue with a much smaller increase.

Also observe that Multi-Topology and Next Shortest perform almost identical for all load values, although Next Shortest seems to always perform slightly better. These methods are the most stable ones, meaning that they perform quite well for all loads.

B. *The Exodus network*

The simulated performance of the Exodus network is depicted in figure 3. Again, notice how Hot Potato deflection performs badly at high loads, and good at low and medium loads.

Also in this network, Next Shortest and Multi-Topology perform about the same, but this time Multi-Topology is mostly the better of the two. Above about 4% loss rate, Multi-Topology also performs better than Hot Potato deflection.

 ISBN:978-1-61208-113-7

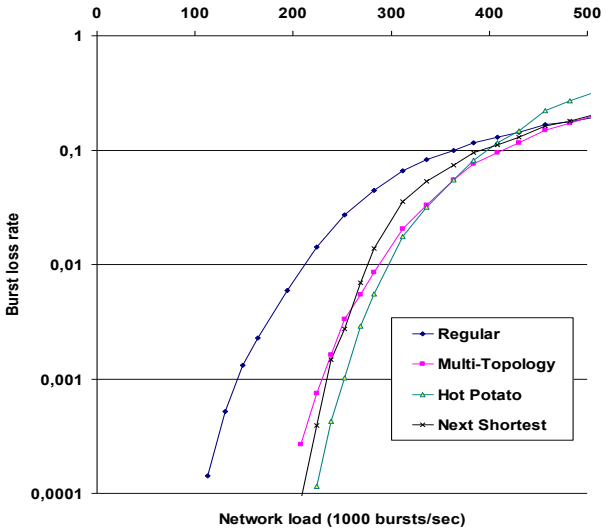


Figure 3. Burst loss rate in the Exodus network with increasing network load.

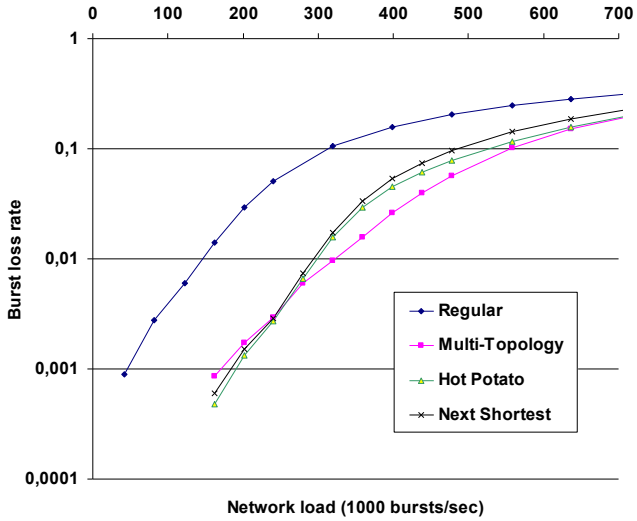


Figure 5. Burst loss rate in the Sprint network with increasing Poisson distributed load.

C. *The Sprint network*

The performance in the Sprint network is seen in figure 4. Here the Hot Potato method is not performing so badly for high loads; in fact it seems that the difference in performance decreases for high loads. In the Sprint network, Multi-Topology deflection is clearly the best method for all load values.

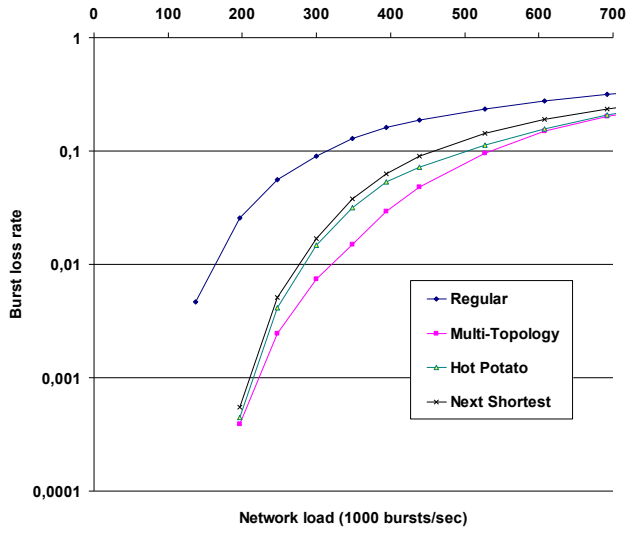


Figure 4. Burst loss rate in the Sprint network with increasing network load.

V. POISSON DISTRIBUTED BURST ARRIVALS

We have also run the experiments described above using Poisson distributed burst arrival processes. Figure 5 shows the burst loss rate in the Sprint network, and figure 6 compares simulated performance of the Multi-Topology method in all the three networks. For low loads, there are only few losses in a second, and then the results are of course less statistical significant.

In our experiments there is not much difference between the performances caused by aggregated self similar traffic and Poisson distributed traffic. First we see this by comparing the plots in the figures 4 and 5. In fact these two plots seem almost identical for burst loss values above 1%. In figure 6 we see that in the Exodus and Sprint networks, Poisson distributed burst loads performs a little better than aggregated self similar traffic, while in the Cost network the situation is reversed. For low loads, however, simulated self similar traffic seems to perform better than Poisson distributed burst load in all three networks, although here we have very little data.

If we assume that a bursty load performs worse (have more burst losses in the network) than a smooth load, and we compare the performance of aggregated self similar traffic (assembled into bursts) with Poisson distributed burst traffic, there is nothing in our experiments that indicates that one of these arrival processes produces smoother burst loads than the other. In future work we will look closer into this problem scenario [10].

 ISBN:978-1-61208-113-7

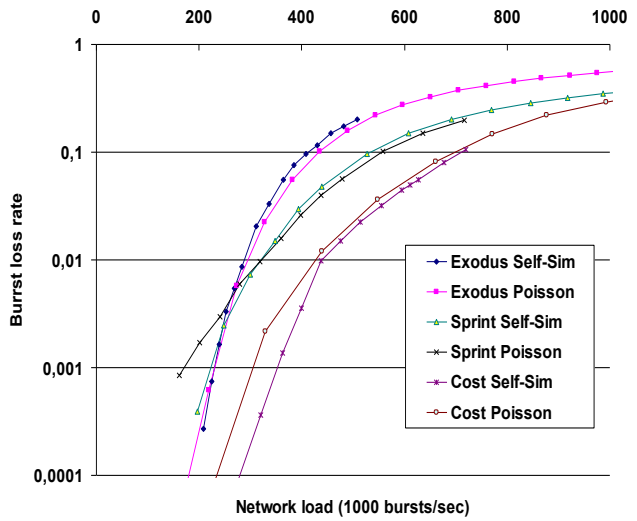


Figure 6. . Simulated performance of Multi-Topology deflection comparing aggregated self similar IP traffic and Poisson distributed bursts in the three networks with increasing network load.

VI. CONCLUSIONS

In this paper we have made a thorough evaluation with realistic traffic of a burst deflection method based upon Multi-Topology (MT) routing. As outlined in the introduction, deflection in OBS networks has been extensively studied before. Also forwarding in sub-graphs of the original optical network has been proposed in order to handle link failures, but, among other things, then the source must know in which sub-graph to forward [19].

MT routing is a novel way to do burst deflection, and hence a way to decrease burst loss probabilities in OBS networks. We have developed a special way to find backup topologies where all links are removed from at least one topology. By pre-calculating shortest paths in these topologies, a burst that can not be scheduled on the primary link, is deflected to an alternative path to the optical egress.

The performance of the MT deflection method is evaluated by comparing it with "Next Shortest" path and "Hot Potato" deflection, as well as with no deflection (just discarding bursts that can not be scheduled on the output link). Three irregular topologies with different characteristics have been used, with number of switches varying from 11 to 45 and ratio between links and switches varying between 1.7 and 2.3.

For high network loads our experiments confirms previous results [5], i.e. that deflection, and in particular Hot Potato deflection, creates more network traffic, and hence makes the burst drop probability higher than Regular routing with immediate dropping of packets when the primary output link is congested.

The arrival processes of bursts have been generated by a hybrid burst assembly method, fed by simulated self similar traffic with variable sized IP packets. At the end of the paper we re-evaluated the performance using a Poisson distributed burst arrival process. The results from these new tests are, for medium and high loads, very similar to the results obtained using self similar IP traffic, and hence strengthen the results from section 4.

Except for very high loads, when Hot Potato routing performs very badly, Multi-Topology seems to be comparable in performance with the two other deflection methods. In the Sprint network, Multi-Topology performs best for all network load values. Next Shortest deflection may (and will in some cases) loop the burst immediately back to the point of congestion, and as long as the original output link is congested, the burst may continue to loop in the network (until discarded when overtaking the control packet). Deflection based on Multi-Topology routing guarantees that such indefinite looping never occurs, and may hence be a viable alternative to other deflection methods in OBS networks.

ACKNOWLEDGEMENTS

Thanks to Audun Fosselie Hansen who took part in the initial design of the OBS simulation model, has provided that backup topologies and programmed most of the algorithm that finds complete sets of backup topologies used in the reported experiments. The author would also like to thank Amund Kvalbein for making the Pareto source module. Our OBS simulation model is based upon another network simulator developed by him and the author. Thanks also to the rest of our colleagues at Simula Research Laboratory.

REFERENCES

[1] Y. Chen, C. Qiao and X.Yu, "Optical Burst Switching (OBS): A New Area in Optical Networking Research," IEEE Network Magazine 18, 16-23 , May/June 2004.

[2] Myungsik Yoo, Chunming Qiao, "Just-Enough-Time (JET): A high speed protocol for bursty traffic in optical networks", In Vertical Cavity Lasers, 1997 Digest of the IEEE/LEOS Meetings, pp. 26–27, Aug. 1997.

[3] J. Ramamirtham, J. Turner, J. Friedman, "Design of Wavelength Converting Switches for Optical Burst Switching", IEEE Journal on Selected Areas in Communications, Vol. 21, No. 7, Sept. 2003.

[4] X. Wang, H. Morikawa, T. Aoyama, "Burst optical deflection routing protocol for wavelength routing WDM networks", Proc. SPIE, 2000.

[5] C.-F. Hsu, T.-L. Liu, N.-F. Huang, "Performance Analysis of Deflection Routing in Optical Burst-Switched Networks", Proceedings IEEE INFOCOM, pp. 66-73, 2003

[6] Zalesky, A.; Hai Le Vu; Rosberg, Z.; Wong, E.W.M.; Zukerman, M.; "Modelling and performance evaluation of optical burst switched networks with deflection routing and wavelength reservation", INFOCOM 2004, Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, Vol. 3, 2004, pp:1864 – 1871.

[7] SuKyoung Lee; Sriram, K.; HyunSook Kim; JooSeok Song, "Contention-Based Limited Deflection Routing Protocol in Optical Burst-Switched Networks", Selected Areas in Communications, IEEE Journal on, Volume 23, Issue 8, Aug. 2005 pp:1596 – 1611

[8] Xiaowei Yang and David Wetherall, "Source Selectable Path Diversity via Routing Deflections", In Proceedings SIGCOMM'06, ACM, September 2006.

[9] P. Psenak et al., “Multi-Topology (MT) Routing in OSPF. RFC 4915, IETF, June 2007

[10] G. Hu, K. Dolzer, C. Gauger, “Does burst assembly really reduce the self-similarity?”, in Optical Fiber Communications Conference, OFC2003, vol.86 of OSA Trends in Optics and Photonics Series, Washington, D. C, 2003, pp.124-126.

[11] Jing Teng, Rouskas, G.N., “Routing path optimiza-tion in optical burst switched networks”, Optical Network Design and Modeling, pp: 1-10, Feb. 7-9, 2005

[12] A. Kvalbein, A. F. Hansen, T. Cicic, S. Gjessing and O. Lysne, “Fast Recovery from Link Failures using Resilient Routing Layers”, In 10th IEEE Symposium on Computers and Communications (ISCC 2005), pp 554-560, IEEE Computer Society, 2005.

[13] A. Kvalbein, A.F. Hansen, T. Cicic, S. Gjessing, O. Lysne, “Fast IP Network Recovery using Multiple Routing Configurations”. In proceedings IEEE 25th Annual Conf. on Computer Communications (INFOCOM) May 2006.

[14] John A. Miller, Andrew F. Seila and Xuewei Xiang, "The JSIM Web-Based Simulation Environment," Future Generation Computer Systems, Vol. 17, No. 2, pp. 119-133. Oct. 2000.

[15] http://www.cs.washington.edu/research/networking/rocketfuel/

[16] Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V., “Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level”, IEEE/ACM Transactions on Networking, Vol. 5, No. 1, pp: 71 – 86 , Feb. 1997.

[17] G. Horn, A. Kvalbein, J. Blomskøld, E. Nilsen, “An Empirical Comparison of Generators for Self Similar Simulated Traffic”, Elsevier Performance Evaluation 64(2): 162-190, 2007.

[18] O’Mahony, M.J., ”Results from the COST 239 project. Ultra-High Capacity Optical Transmission Networks”, 22nd European Conference on Optical Communication, pp: 15-19, Sept. 1996

[19] M. T. Frederick, P. Datta, A. K. Somani “Sub-graph routing: A generalized fault-tolerant strategy for link failures in WDM optical networks”, Computer Networks, Vol. 50, No. 2, Feb. 2006, pp. 181-199, Elsevier 2006.

[20] S. Gjessing, “A novel method for re-routing in OBS networks”, In International Symposium on Communications and Information Technologies, 2007. ISCIT '07. Sydney, NSW, October 2007.

 ISBN:978-1-61208-113-7

Modeling and Evaluation of SWAP Scheduling Policy Under Varying Job Size Distributions

Idris A. Rai
Makerere University
Faculty of Computing and Informatics Technology
Kampala, Uganda
rai@cit.mak.ac.ug

Michael Okopa
Makerere University,
Faculty of Computing and Informatics Technology
Kampala, Uganda
michaelokopa@yahoo.co.uk

***Abstract*—Size-based scheduling policies have been shown to be effective resource allocation policies in computing and networked environments. One of the recently proposed size-based scheduling policy is called SWAP. It is a non-preemptive, threshold based policy that was proposed to approximate the Shortest Job First (SJF) policy by introducing service differentiation between short and large jobs such that short jobs are given service priority over the large jobs. Original study of the SWAP scheduling policy was based on only simulations, which are known to have a number of restrictions. In this paper, we derive SWAP models and evaluate the scheduling policy using workloads that have varying distributions. In contrast to simulations, the models enable fast analysis of the scheduling policy under a wide range of input parameters. Numerical results obtained from the derived models show that SWAP approximates SJF better for heavy-tailed workloads than for exponentially distributed workloads. We also show that SWAP performs significantly better than First Come First Serve (FCFS) and Processor Sharing (PS) policies regardless of the distribution of the workload.**

***Keywords*-Size-based scheduling; conditional mean response time; temporal dependence.**

I. INTRODUCTION

Motivated by the persistent evidence of heavy-tail distribution of stored and transferred file sizes, size-based scheduling policies have been widely studied for efficient resource allocation in time-sharing computing environments. Most size-based scheduling policies however require the knowledge of job sizes, which present a major limitation to their practical implementations. Examples of common size-based scheduling policies include Shortest Job First (SJF), which is a non-preemptive scheduling policy that gives service to the shortest job immediately after the job in service completes. Shortest Remaining Processing Time (SRPT), on the other hand, is a preemptive variant of SJF. It favors short jobs by giving service to the job in the queue that has the shortest remaining processing time. In order to know the remaining processing time however, one needs to know the total service required by the job (i.e., the size of the job). SRPT is known to be the optimal policy in terms of providing the minimum mean response time [3], [1]. Despite its optimum performance, SRPT scheduling is not widely used in practice particularly in network environments due to lack of information on flow sizes [2]. Some authors have as a result investigated the use of SRPT without accurate knowledge of job sizes [4], [5].

Blind scheduling policies, which are policies that don't require job sizes, are therefore often preferred in practice due to their implementation simplicity. Some popular examples of blind scheduling policies include Processor Sharing (PS), First Come First Served (FCFS), and Round Robin [3]. There also exist blind size-based scheduling policies; which are blind scheduling policies that take into account a notion of size. The most popular example is Least Attained Service (LAS) first, which is a preemptive scheduling policy that favors short jobs by giving service to the job in the system that has received the least service [3], [7], [8]. LAS doesnt require the knowledge of job sizes, and therefore can be used in network routers. However, its implementation requires a router to keep track of states of each flow that traverses the router, which may be a very daunting task in high speed networks.

In this paper, we study another recently proposed size-based scheduling policy called SWAP [6]. SWAP is size-based policy that was proposed to improve on the response time of short jobs while reducing the overhead required in identifying job sizes. The basic idea behind SWAP is to use the measured serial correlation of the service times to estimate the missing information of job sizes. Once these reliable estimates of the job service times are available, large jobs are delayed by putting them at the tail of the queue whereas short jobs are kept at the head of the queue for immediate service. SWAP uses a threshold value to decide which jobs are to be delayed. In such a way, delayed long jobs are served after most short jobs in the queue have completed their service. We review SWAP policy in detail in Section III-A.

In an effort to improve the performance of systems, studies have shown that the recent past is indicative of the near future. It is a generalization of the idea of locality which can be exploited to leverage the need for knowing sizes of all jobs. Furthermore, it can be observed that the real workload data is far from independently and identically distributed. Instead, similar jobs tend to arrive within bursty periods. This observation is vividly true when the workload exhibits heavy tailed distribution. SWAP was originally proposed to take advantage of this observation in accomplishing its goals. Authors in [6] proposed SWAP to approximate the behavior of the optimal SJF scheduling policy by using workload temporal dependence to forecast job service times without any a priori

 ISBN:978-1-61208-113-7

knowledge of upcoming job demands.

Original study of SWAP was conducted through *only* simulations of the policy. Simulations techniques however have some limitations; they are restrictive in the sense that often only programmers can comfortably adopt them, and it often takes long to obtain results for a wide range of input parameters. To complement the original work on SWAP, in this paper, we derive analytical models of SWAP in terms of the conditional mean response time of jobs. The models can be used for quick performance evaluation of the policies. We investigate the performance of SWAP in terms of approximating SJF under varying job size distributions, and numerically compare the performance of SWAP to FCFS and Processor sharing (PS) policies.

The rest of the paper is organized as follows: in the next section, we present mathematical background that guides models derivation of SWAP policy. In Section III, we review SWAP scheduling and derive its models. We evaluate SWAP in Section IV, and finally conclude the paper in Section V.

II. Mathematical Background

Let's denote the probability density function (pdf) of a job size distribution as $f(x)$. The cumulative distribution function ($F(x)$) is obtained as $F(x) = \int_0^x f(t)dt$, and the survival function (or reliability function) is given as $F^c(x) = 1-F(x)$. We define $\overline{x_x^n} = \int_0^x t^n f(t)dt$ to be the n^{th} moment for jobs that are less than or equal to x. Therefore, $\overline{x_x}$ is the mean and $\overline{x_x^2}$ is the second moment of the job size distribution due to job sizes less than or equal to x. The mean and second moments are obtained when the value of subscript x is infinity.

Let x_l be a large job under SWAP scheduling, i.e., any job that is greater than a specified threshold (x_t). It follows that $\overline{x_{x_l}^n} = \int_{x_l}^\infty t^n f(t)dt$ is the n^{th} moment for job sizes greater than x_l.

The load due to jobs with sizes less than or equal to x is given as $\rho_x = \lambda \int_0^x tf(t)dt$ while the load due to jobs with sizes greater than x_l is given as $\rho_{x_l} = \lambda \int_{x_l}^\infty tf(t)dt$. Also $\rho_{x_l} = \rho - \rho_{x_l}$ where ρ is the total load in the system. We next define the expressions for the conditional mean response time under FCFS and SJF, which we shall use when deriving the models for SWAP policy.

An arriving job to a FCFS queue has to wait for all jobs it finds in the queue upon its arrival. Therefore, the conditional average response time of a job of size x in an M/G/1/FCFS system is given as

$$T(x) = x + W(x), \tag{1}$$

where $W(x) = \frac{\lambda \overline{x^2}}{2(1-\rho)}$ is the mean waiting time due to jobs in the system. Assume that an arriving job x finds only the jobs that are less than or equal to a job size x_t in the M/G/1/FCFS queue. Its conditional average response time $T(x)$ is given as

$$T(x_t) = x + W(x_t), \tag{2}$$

where $W(x_t) = \frac{\lambda \overline{x_{x_t}^2}}{2(1-\rho_{x_t})}$.

Under SJF, the shortest job in the queue is given non-preemptive priority. Thus, at every completion instant of a job in the server, the next job to receive service is the smallest job in the queue. A job of size x is therefore delayed by only jobs in the system that are less than or equal to its size. The conditional average response time of the job size of x under SJF is given as

$$T(x_x) = x + W(x_x), \forall x > 0, \tag{3}$$

where $W(x_x) = \frac{\lambda \overline{x_x^2}}{2(1-\rho_x)^2}$.

We will numerically evaluate the SWAP models under job sizes with exponential distribution and job sizes with Bounded Pareto distribution to mitigate workloads with varying variances. The variability of a job size distribution is determined by its Coefficient of Variation (C), which is defined as the ratio of the standard deviation to the mean of a distribution. Exponential distribution has a low variability since its $C = 1$, whereas a Bounded Pareto distribution has a high variability ($C > 1$). In this paper, we specifically use exponential distribution and Bounded Pareto $BP(10, 5 * 10^5, 1.1)$ distributions with mean values of 72.7 to numerically evaluate SWAP models. Similar distributions have been used in [1], [7]. The probability density function of an exponential distribution is given as:

$$f(x) = \mu e^{-\mu x}, x \geq 0, \mu \geq 0. \tag{4}$$

Bounded Pareto distributions have commonly been used to evaluate the performance of systems under heavy tailed workloads with high variance [1], [7], [8]. In contrast to Pareto distributions which assume infinite largest job size, Bounded Pareto distributions can be used to represent realistic workload with known largest values. We denote Bounded Pareto distribution by $BP(k, P, \alpha)$ where k and P are the minimum and the maximum job sizes and α is the exponent of the power law. The pdf of the Pareto is given as:

$$f(x) = \frac{\alpha k^\alpha}{1-(k/P)^\alpha} x^{-\alpha-1}, \quad k \leq x \leq P, \quad 0 \leq \alpha \leq 2. \tag{5}$$

In the next section, we discuss SWAP scheduling and derive its models.

III. SWAP Scheduling Models

A. A review of SWAP Scheduling

SWAP is a class-based, non-preemptive, size-based scheduling policy where jobs are classified into two classes based on their sizes, namely *short* (x_s) jobs and *large* (x_l) jobs classes. SWAP uses a rather naive definition of job size based on threshold (x_t), which can be dynamic. All jobs that are less than or equal to x_t are classified as short whereas jobs that larger than x_t are classified as large jobs. The main goal of SWAP scheduling policy is to approximate the SJF scheduling policy so as to favor short jobs without apriori knowledge of job service requirements or job sizes. It reduces the mean response time for short jobs by designating higher priority to short jobs compared to large jobs.

 ISBN:978-1-61208-113-7

SWAP starts by serving all arriving jobs to a queue in FCFS manner, and compares the service given to each job to the threshold value. If a large job is served, next the entire queue is *scanned* whereby size of each job in the queue is computed and jobs in the queue are classified and marked as large or short. Once classified, short jobs are moved at the head of the queue and receive service before large jobs that are kept at the tail of the queue. We shall term these scanned large jobs as (*delayed*). Jobs within a class are serviced in their order of arrival using FCFS scheduling. Once a job has been classified under SWAP scheduling it will belong to that class for all duration of its stay in the queue.

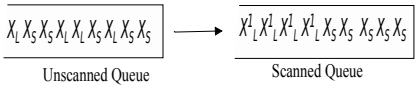


Fig. 1. Illustration of unscanned and scanned queues in SWAP(1)

It can be seen that compared to FCFS scheduling, SWAP favors short jobs to the expense of delaying large ones within the queue. Intuitively, SWAP policy should provide significant performance improvement in terms of reducing mean response time of short flows particularly for highly varying workloads where only a tiny fraction of jobs contribute to more than half of the system load.

Jobs that arrive to a scanned SWAP queue are buffered in order of their arrival and are eventually served in FCFS order once all scanned jobs in the queue have completed their service. We call the state where served jobs in the queue are not scanned an *unscanned* state of the queue. SWAP therefore alternates between scanned and unscanned states.

Large jobs under SWAP scheduling can be delayed more than once. We denote SWAP(i) as SWAP scheduling policy which delays large jobs i times. Consider SWAP(1), the service time of a large job that arrives to an unscanned queue is interrupted by the short jobs that were in the queue upon its arrival and the one large job that triggers the scanning of the queue. The scanning evet will delay the large jobs once. They will receive service immediately after the short scanned jobs have completed their service. It can be seen that under SWAP(1), the newly arriving short jobs in a scanned queue have to wait until all scanned large jobs in the queue complete their service before they can receive any service. Figure 1 illustrates SWAP(1) scheduling at scanned and unscanned states where X_S and X_L^1 denote a short job and large job that has been delayed 1 time.

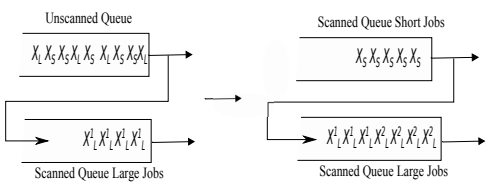


Fig. 2. Illustration of SWAP(2)

Increasing delays of large jobs under SWAP improves the service of short jobs further by avoiding disruptions of their service due to some large jobs that arrived in the queue before them. Consider SWAP(2), for instance, once the first short scanned jobs are served, the large scanned jobs wont receive service immediately. Instead, the server will serve the arriving unscanned jobs in the queue until the scanning is triggered again by an unscanned large job. At that point, the entire queue will be scanned to classify the jobs. The server will then give priority to the scanned short jobs in the queue before it can serve the large jobs that have been delayed twice.

Once large jobs have been delayed 2 times under SWAP(2), they receive service immediately after all classified short jobs in the queue have completed their services. Figure 2 demonstrates SWAP(2), where again X_L and X_S denote large and small jobs, and X_L^i are large jobs that have been delayed i times. Note that we represent SWAP with two queues where the second queue hosts delayed large jobs after they are scanned. Delayed scanned jobs wait in the second queue until they are delayed for i times equivalent to $i-1$ new scanning events since the time they were scanned.

B. Modeling SWAP Scheduling

In this section, we derive models of SWAP scheduling in terms of conditional mean response time of short jobs and large jobs. We consider a tagged job that arrives to an M/G/1/SWAP(i) system at scanned and unscanned states separately.

1) Models for arrivals at unscanned state: Assume a tagged job arriving to a SWAP queue in an unscanned state, but *just* before a large job receives service. Recall that such a large job will trigger the scanning of the queue. If the tagged job is short it will be placed at the tail of the short jobs class, otherwise it will be placed at the tail of the queue.

Let's consider SWAP(1), the tagged short job will be delayed by the mean residual life of the large job that triggered the scanning and will then wait for all scanned short jobs it finds in the queue before it receives service. The waiting time of the tagged job due to these jobs is given by $W(x_t)$ defined in Section II. On the other hand, the tagged large job will be delayed by all jobs it finds in the system upon its arrival by the mean waiting time denoted by $W(x)$. Note that $W(x)$ represents the delay of short and large scanned jobs. The corresponding delays due to separate jobs classes are denoted as $(W(x_l))$ for large jobs and $W(x_t)$ for short jobs. The resulting conditional mean response times under SWAP(1) at unscanned state are given as:

$$T(x) = \begin{cases} W(x_t) + x & x \leq x_t \\ W(x_l) + W(x_t) + x & x > x_t \end{cases}$$

Note that the conditional mean response time for a large job here is the same as the conditional mean response time of the job under the FCFS queue (Equation 1).

We now derive the models for SWAP(2), under which large jobs are delayed twice before they receive service. We assume the steady state case of the queueing system at which arriving jobs will always find scanned large jobs that have been delayed

once waiting in the queue. These delayed jobs will receive service immediately after the newly scanned short jobs have completed their service.

Again, assume a tagged job as the last job that arrives to the queue before the queue turns from unscanned state to scanned state. Let the tagged job be a short job; it will be delayed by all short jobs it finds in the queue plus the remaining service of the large job it finds in the server when it arrived. The expression of its conditional mean response time is the same as the conditional mean response time of short job arriving at SWAP(1) at unscanned state shown in Equation (III-B1). If the tagged job is a large job, it will be delayed by all jobs it finds in the queue, which include all the short and large jobs that have just been scanned, and the large jobs that were scanned in the previous scanning event and the large jobs that were delayed in the queue upon its arrival. Since itself has to be delayed twice, the job will also be delayed by arriving new short jobs, both unscanned and scanned. The conditional mean response time of the tagged job arriving during unscanned state of SWAP (2) is therefore given as follows:

$$T(x) = \begin{cases} W(x_t) + x, & x \leq x_t \\ 2W(x_l) + 2W(x_t) + \overline{x}F^c(x_l) + x, & x > x_t \end{cases} \quad (6)$$

The general expression for SWAP(i) can be obtained using iterative method, and is given in Equation (7).

$$T(x) = \begin{cases} W(x_t) + x, x \leq x_t \\ iW(x_l) + iW(x_t) + (i-1)\overline{x}F^c(x_l) + x, \\ \qquad\qquad x > x_t \end{cases} \quad (7)$$

Observe that the derived expressions for conditional mean response times dont show performance gain acquired by short jobs from delaying large jobs more times. To intuitively see the reduction on short jobs response times, note that if large jobs arent delayed, the short jobs would be served after the scanned large jobs which would in turn increase their mean response time by $W(x_l)$.

2) Models for arrivals at scanned state: We now consider a tagged job arriving to a scanned queue. We derive models for the worst case scenario where the tagged job finds in the queue scanned short jobs being serviced, all scanned large jobs and other unscanned jobs including at least one large unscanned job waiting in the queue. This tagged job will experience the longest mean response time under SWAP. Other scenarios that we shall skip due to space limitation include the tagged job arriving just after a scanning event and a tagged job arriving to a scanned queue with only short unscanned jobs. The analyses of these skipped scenarios however are straight forward.

Let's assume the tagged short job is arriving to a scanned queue of SWAP(1) policy under the worst case scenario presented above, its service will be delayed by all scanned jobs it finds in the queue, all unscanned short jobs that it finds in the queue, and finally one large unscanned job that will trigger the next scan. The conditional mean response time of the tagged short job will include mean waiting time due to the service of the remaining scanned short jobs ($W_r(x,t)$), mean service time of scanned large jobs ($W(x_l)$), the service of the single large job that triggers the scanning ($\overline{x}F^c(x_l)$), and finally the mean service time of unscanned short jobs it finds in the queue ($W(x_t)$). If we assume that the tagged job arrives at the queue at a random point after the queue is scanned, we can further approximate $W_r(x_t)$ as $W(x_t)/2$.

The tagged large job, on the other hand, will additionally be delayed by large jobs that were newly scanned along with itself by a mean waiting time of $W(x_l)$. The conditional mean response time for the tagged job is obtained as follows:

$$T(x) = \begin{cases} W(x_l) + 3W(x_t)/2 + x + \overline{x}F^c(x_l), & x \leq x_t \\ 2W(x_l) + 3W(x_t)/2 + x + \overline{x}F^c(x_l), & x > x_t \end{cases}$$

For SWAP(2), the tagged short job arriving after scanning will see in the queue scanned short jobs, scanned large jobs that have been delayed once, and unscanned short jobs and large jobs. In contrast to a short job under SWAP(1), the service of the tagged short job here will not be interrupted by the delayed large jobs since they have to be delayed once more. Therefore, the job's response time will be due to the remaining scanned short jobs by approximate of $W(x_t)/2$, unscanned short jobs it finds in the queue by mean waiting time of $W(x_t)$, and the one unscanned large jobs that will trigger the next scanning event.

For a tagged large job, it will be additionally delayed by the scanned large jobs it finds in the queue by mean waiting time of $W(x_l)$, all short jobs that will arrive until the next scanning event by mean waiting time of $W(x_t)$, the large job that will trigger the next scanning event, and any large jobs that it finds in the queue by their mean waiting delay of $W(x_l)$. Using similar arguments as before, we obtain the expression for conditional response time for the job as follows:

$$T(x) = \begin{cases} 3W(x_t)/2 + x + \overline{x}F^c(x_l), & x \leq x_t \\ 2W(x_l) + 5W(x_t)/2 + x + 2\overline{x}F^c(x_l), & x > x_t \end{cases}$$

The general expression for SWAP(i) model for short jobs isnt very obvious to derive. We present instead the general model large jobs arriving at scanned states as follows:

$$T(x) = iW(x_l) + (i+1/2)W(x_t) + x + i\overline{x}F^c(x_l), i \geq 1, x > x_t$$

In the next section, we present numerical results showing the performance of SWAP and its comparison with SJF and PS scheduling policies

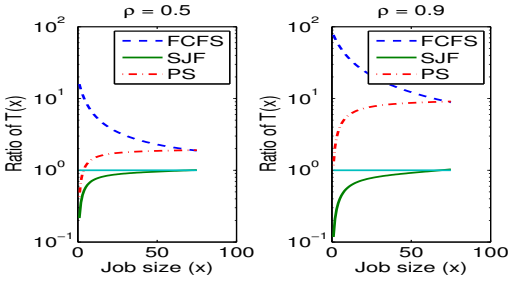


Fig. 3. Ratios of $T(x)$ vs x exponential workloads, $x_t = 75$

IV. Performance Evaluation

In this section, we use the derived SWAP models to evaluate its performance. We look at how SWAP approximates the SJF for short and large jobs, and we also compare its performance with that of FCFS and PS policies. We also investigate the impact of threshold values (x_t) to the performance of SWAP. Processor sharing (PS) is one of the mostly studied policy in time-sharing operating systems. It is also known as a fair scheduling policy providing conditional mean response time of $x/(1-\rho)$ for a job with size x.

We use exponential and Bounded Pareto distributions presented in Section II at low and high system load values of $\rho = 0.5$ and $\rho = 0.9$ respectively. For each set of result, the threshold values were chosen such that ρ_{x_t} under both considered distributions are close to each other. Due to space limitations, we numerically evaluate SWAP models for jobs that arrive to unscanned state only which we derived in Section III-B1.

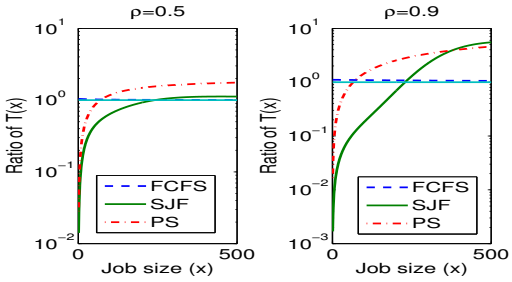


Fig. 4. Ratios of $T(x)$ vs x for exponential workloads, $x_t = 500$

Figures 3 and 4 show the ratios of conditional mean response time of short jobs under FCFS, PS, and SJF policies to that of SWAP for threshold values of $x_t = 75$ and $x_t = 500$ respectively for the case of exponentially distributed workload. It can be observed from the figures that SWAP indeed approximates the SJF for low load and short threshold values where the response time ratio of SWAP to SJF is always close to one. The approximation is more accurate for larger jobs compared to short jobs. The estimates under SWAP are less accurate for larger thresholds and higher loads as seen in Figure 4 at $\rho = 0.9$.

We can also see from the figures that SWAP performs much better than FCFS and PS for small threshold values regardless of the load. At high threshold values, SWAP offers similar mean response time as FCFS since the scanning event is triggered by very large jobs which makes SWAP scheduling for even short jobs the same as SJF. This can be seen in Figure 4 where $x_t = 500$ and $\rho_{500} = 0.89$.

Similarly, Figures 5 and 6 show results of SWAP in comparison with SJF, FCFS, and PS under Bounded Pareto distributed workloads presented in Section II. Compared to the results for exponential workloads, we can quickly see that SWAP approximates SJF better under Bounded Pareto distribution for all load values and varying threshold values. Bounded Pareto workloads are heavy tailed meaning more than 99% of their

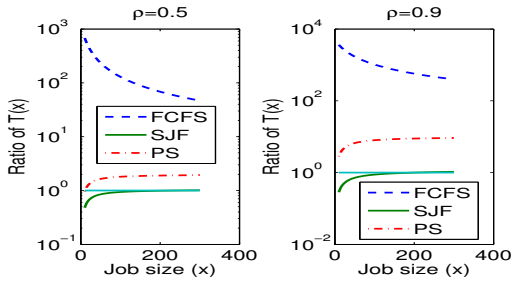


Fig. 5. Ratios of $T(x)$ vs x for Bounded Pareto workloads, $x_t = 300$

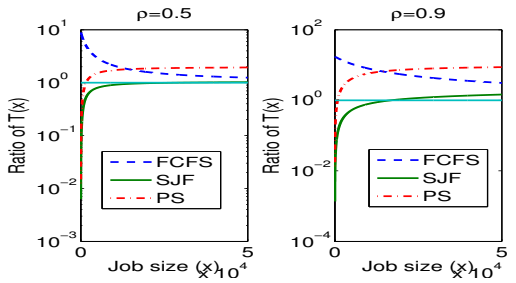


Fig. 6. Ratios of $T(x)$ vs x for Bounded Pareto workloads, $x_t = 50000$

jobs are short and constitute to only half of the total load. Consequently, most of the jobs under Bounded Pareto workloads are classified as short, and by delaying the few very large ones, the performance of short jobs is improved significantly. However, at large thresholds, regardless of system load, very short jobs under SWAP experience longer mean response times than under SJF (see Fig. 6), meaning SWAP is inaccurate in approximating SJF.

We also observe that the mean response time provided by SWAP compared to that of FCFS and PS under Bounded Pareto job size distributions follow similar trends to the performance under exponential distribution. In general, SWAP performs much better than FCFS for short jobs. SWAP also performs better than PS except for very few short jobs under SWAP with large threshold values such as in Fig. 6.

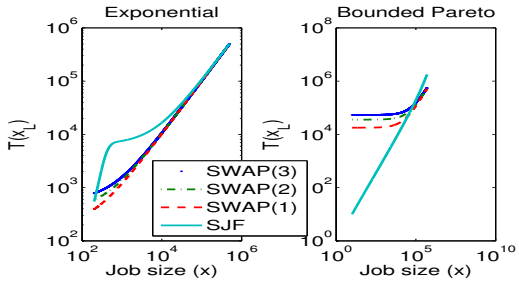


Fig. 7. $T(x)$ vs x_L for exponential with $x_t = 200$ and Bounded Pareto with $x_t = 5000$, $\rho = 0.9$

We finally investigate the conditional mean response time of large jobs under exponential and heavy tailed workloads for SWAP system with high and low loads. We specifically compare the performance of large jobs under SWAP and SJF to see how well SWAP approximates SJF for large jobs, and

 ISBN:978-1-61208-113-7

under SWAP(i) for $i = 1, 2, 3$ to see how much additional response time large jobs experience as a result of being delayed further. We exclude results that compare SWAP to PS and FCFS for large jobs due to space limitation.

We observe from Figure 7 that for the considered parameters, all SWAP(i) offers similar performance and all approximate SJF well for the very large jobs under exponential workload. We also observe that some shorter jobs experience longer mean response time under SJF than under SWAP policies. The situation is however very different for heavy tailed workloads where SWAP with larger i offering noticeably worse conditional mean response time than SWAP with smaller i. The performance of SWAP also significantly differs from SJF specifically for shorter jobs where SWAP performs worse and for the largest jobs where SWAP instead performs better that SJF. The largest jobs under SJF are interrupted by all jobs in the system which is not the case for SWAP. On the other hand, shorter jobs under SWAP are interrupted by large jobs of any size compared to only jobs that are less than their sizes under SJF.

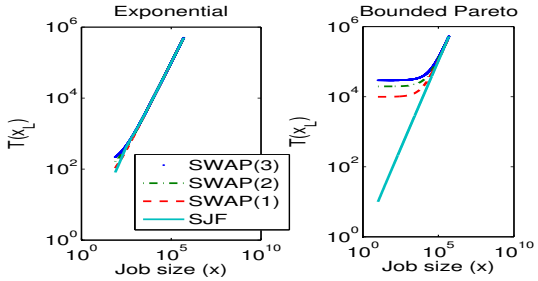


Fig. 8. $T(x)$ vs x_L for exponential with $x_t = 75$ and Bounded Pareto with $x_t = 300$, $\rho = 0.5$

Figure 8 shows the results at low load of $\rho = 0.5$, and smaller threshold values of $x_t = 75$ and $x_t = 300$ for exponential and heavy tailed workloads respectively. We can easily see that the performance of all SWAP(i) policies is similar for the case of exponential distribution. In this case, we also see that SWAP approximates SJF very well. However, for heavy-tailed workloads the performance of shorter jobs is still worse under SWAP than under SJF but is just slightly less than for the case of $\rho = 0.9$. The performance of the largest jobs however is the same under all policies showing better approximation of SJF by SWAP for the largest jobs compared to the case for high load and large threshold shown in Figure 7.

We conclude therefore that the performance of SWAP in terms of jobs mean response time depends on the distribution of workloads. Similarly, its accuracy in approximating SJF also depends on the distribution of the workloads. In particular, SWAP performs well under both respects for short jobs that arrive to unscanned queue for heavy-tailed workloads. In contrast, SWAP performs poorly in terms of offering higher mean response time than SJF to large jobs just above the threshold value regardless the load and threshold value. SWAP scheduling is also inaccurate in approximating SJF for large jobs except for low load and shorter thresholds.

V. Conclusion

We modelled and evaluated the recently proposed SWAP scheduling policy under varying workload distributions. The numerical results that we obtained from the derived models show that SWAP can be used to approximate Shortest Job First (SJF) however it is more accurate for workloads with highly varying job sizes such heavy-tailed job size distributions. This is because heavy-tailed distributions exhibit expectation paradox that is a clear manifestation of temporal dependence, which is the basic assumption under which SWAP was proposed. The comparison of SWAP with FCFS and PS also show that SWAP is a more superior policy in terms of reducing the mean response time of short jobs. We further observed that in contrast to exponentially distributed workloads, especially at high threshold and load values, large jobs under SWAP suffer very negligible penalty.

In this paper, we presented numerical results of SWAP at unscanned state only due to space limitations. In the future, we will numerically investigate SWAP at scanned state as well. We will also explore the use of SWAP for networked environments with Internet flows as transferred entities. In contrast to jobs, flows don't arrive at a system all at once, making it very hard to immediately infer their sizes. We will check to see if per-connection buffer occupancy will approximate SWAP in such environments. Finally, we intend to validate the SWAP models derived in this paper using simulations of the policies.

Acknowledgement

This work was partially funded by CISCO University Research Fund, a corporate adviced fund of Silicon Valley Community Foundation.

References

[1] N. Bansal and M. Harchol-Balter. Analysis of SRPT Scheduling: Investigating Unfairness. In *Sigmetrics 2001 / Performance* 2001, pp. 279-290, June 2001.

[2] M. Harchol-Balter, M. Schroeder, B. Bansal, and M. Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems (TOCS)* 21, 2 (May 2003).

[3] Leonard Kleinrock, *Queueing Systems*, Volume II. Computer Applications: John Wiley & Sons, 1976.

[4] D. Lu, P. Dinda, Y. Qiao, H. Sheng, F. Bustamante. Applications of SRPT Scheduling with Inaccurate Scheduling Information. In *Proceedings of IEEE MASCOTS*, October (2004).

[5] D. Lu, H. Sheng, and P. Dinda. Size-based scheduling policies with inaccurate scheduling information.In *Proceedings of IEEE MASCOTS* (2004).

[6] N. Mi, G. Casale, and E. Smirni. Scheduling for Performance and Availability in Systems with Temporal Dependent Workloads. In *The International Conference on Dependable Systems and Networks*, pp. 336-345, 2008.

[7] I. A. Rai, G. Urvoy-Keller, and E. Biersack. Analysis of LAS Scheduling for Job Size Distributions with High Variance. In *Proc. of ACM SIGMETRICS'03*, pp. 218-228, June 2003.

[8] I. A. Rai, G. Urvoy-Keller, M. K. Vernon and E. W. Biersack. Performance Analysis of LAS-based Scheduling Disciplines in a Packet Switched Network. In *Proc. ACM SIGMETRICS*, June 2004.

 ISBN:978-1-61208-113-7

Optical Protection with Pre-configured Backup Paths and Limited Backup Resource Sharing

Krishanthmohan Ratnam, Mohan Gurusamy, and Kee Chaing Chua
Department of Electrical and Computer Engineering, National University of Singapore
Email: {eleratn,elegm,eleckc}@nus.edu.sg

***Abstract*—In this paper, we consider provisioning protection in WDM optical networks with pre-configured backup paths. In the traditional protection approach, backup resources are not shared among pre-configured backup paths, and thus resources are not utilized efficiently. We propose a protection approach with the use of a switch architecture, which allows limited sharing of backup resources among pre-configured backup paths (referred to as *pre-configured backup protection with limited sharing (PBPLS)*). The architecture uses switching components with a flexible feature of splitting optical power on need basis in addition to directing the power towards one output port only. Further, configuration can be done which connects two (or more) input ports to the same output port at the same time. These features allow sharing backup resources while provisioning pre-configured backup paths. This approach can be adopted in networks in small geographical area such as metro networks since the power splitting feature is used. While sharing backup resources in this approach, we consider power loss particularly due to potential repeated power splitting. Amplifiers can be used, at additional cost, to compensate the power loss. Instead, we adopt an approach of limiting the number of power splitting to small values to reduce the power loss. Constraining the number of power splitting limits the degree of backup sharing. Through simulation experiments in a single class and multi-class traffic scenarios, we demonstrate that, even with the small number of power splitting such as one or two, significant improvement in blocking performance can be achieved.**

***Keywords*-optical networks; wavelength-division-multiplexing; survivability;**

I. INTRODUCTION

Survivability or fault tolerance is an important requirement in wavelength-division-multiplexing (WDM) optical networks. Among the several survivability approaches, provisioning optical layer protection with pre-configured backup paths such as optical dedicated protection (or 1:1 protection) is preferred for traffic which require short recovery time. In this approach, a backup path is configured at the time when the connection is established. In the event of a component failure on a primary path, this approach requires no further switch configuration to set up the backup path. This protection approach has been investigated in research works under several scenarios such as path, segment, and link based protection, protection with traffic grooming, differentiated survivability services, and protection with multi-line-rate consideration. In [1], two 1:1 path protection methods, static and dynamic have been investigated. The static method provides fixed primary and backup paths, and the dynamic method allows rearrangement of backup paths. The work in [2] investigates capacity utilization and protection switching time for a dedicated path protection scheme and protection approaches which share backup resources. Dedicated protection for traffic grooming of sub-lambda traffic using a generic grooming-node architecture has been investigated in [3]. In [4], a comparison of schemes which include path and segment based protection for differentiated availability-guaranteed services is given. The recent work in [5] investigates dedicated protection approaches considering various transmission rates of wavelength channels.

A major drawback in provisioning pre-configured backup paths using the traditional protection approach is its inefficient resource usage. Unlike the optical layer shared protection approach, in this approach backup resources are not shared among the pre-configured backup paths and thus resources are not utilized efficiently. The traditional optical shared protection has long recovery time, in which backup paths are not pre-configured and backup wavelength links can be shared by other backup paths. The work in [2] shows that, with 10ms switch configuration time, the recovery times of dedicated and shared protection approaches are 3ms and 56ms respectively under a distributed protocol (for a random demand of 30 connections on a representative network topology). The configuration time of switches widely used could be several 10's of ms and the difference in recovery time for the two approaches would, therefore, be even more significant. Several mission critical applications require short recovery time. Pre-configured backup protection is suitable for such applications. The shared protection approach may not satisfy their stringent recovery time needs.

We propose a protection approach which allows limited sharing of backup resources among pre-configured backup paths (referred to as *pre-configured backup protection with limited sharing (PBPLS)*). The proposed approach can be used under single component failure scenarios. To allow such resource sharing, we use the switch architecture proposed in [6]. The architecture has the following flexible features. In addition to directing the input power towards one output port only (like the traditional switches), the power can be split on a desired sub-set of output ports on need basis. When the switch is pre-configured to split power on two output ports, the traffic can be switched on one of the ports with the split power which requires no further configuration. Further, the switch can be configured to connect two (or more) input ports to the same output port at the same time. With this pre-configuration, the traffic can be switched from one of the input ports to the

 ISBN:978-1-61208-113-7

same output port which requires no further configuration. In the proposed approach, when backup paths are set up, similar pre-configurations can be done so that backup resources can be shared. The recovery time in this case is equivalent to the case of the traditional dedicated protection approach since no further configuration is needed at intermediate nodes. This protection approach can be adopted in networks in small geographical area such as metro networks since the power splitting feature is used when sharing backup resources.

In the proposed approach, we consider limited sharing of backup resources. This is because of power loss when power splitting is used for backup sharing. Particularly, when repeated or cascading of power splitting occurs, power will be reduced significantly. One solution is to compensate power using amplifiers at additional cost. In this paper, we adopt the approach of limiting the number of power splitting on a backup path to reduce the power loss. Constraining the number of power splitting limits the degree of backup sharing. We investigate for small values for the maximum number of power splitting (one to three).

As explained above, the proposed protection approach utilizes the flexible features of the switch architecture used in this paper. The widely used traditional optical switches such as MEMS switches [7] do not support these features because of architectural limitations, and therefore similar protection approach cannot be adopted. The proposed approach can be employed in broadcast-and-select based architectures which are widely considered in optical burst/packet switching networks [8] [9]. These architectures generally consist splitters and semiconductor optical amplifiers (SOAs). Since SOAs are used, power loss due to power splitting would be compensated and the need for limiting the number of power splitting may not arise (or reduced). However, we do not use these architectures in this paper because of their high power loss and high cost. Splitters used in these architectures always split power towards all the output ports and thus significantly a large amount of power is wasted. Further, these architectures are expensive since a large number of SOAs are required. The switch architecture used in this paper uses components with the flexibility of controlled power directing and splitting as explained above. Therefore, it reduces power wastage significantly. In addition to this, we do not use amplifiers in the architecture in order to reduce the cost.

In [10], an approach has been proposed to improve resource usage in which a pre-configured backup path can share resources of non pre-configured backup paths. Unlike this approach, this paper investigates sharing backup resources among pre-configured backup paths. Power splitting has been considered in [11] [12] when provisioning protection. In [11], a 1+1 dedicated protection approach (traffic is simultaneously sent via the two alternate paths) has been investigated in which splitters are used in broadcast-and-select OADMs for a ring topology network. This work does not consider backup sharing. In [12], splitters are used in tree-based protection for multicast traffic. In this work, backup sharing is considered and nodes may require reconfigurations in the event of a failure. In our work, we consider unicast only and the proposed approach avoids reconfigurations when failure occurs. In the following sections, we illustrate the switch architecture first, and then illustrate our proposed protection approach.

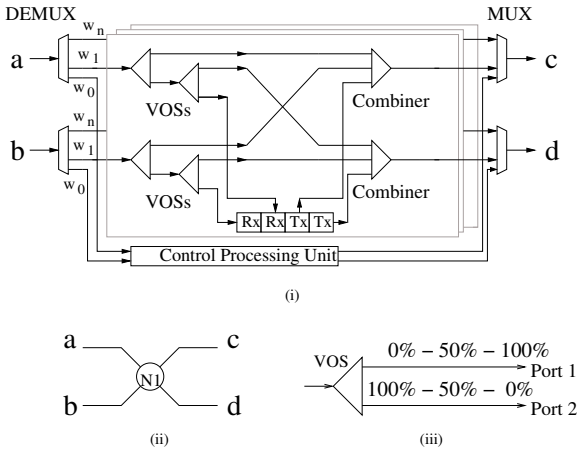


Fig. 1. Switch architecture with two input and two output links

II. Optical switch architecture

The proposed switch architecture [6] is shown in Fig. 1(i). In [6], we have investigated the transmission of bursty traffic. The switch can be used for the transmission of circuit level traffic also (transmission through lightpaths). We consider the switch for a node with two input and two output links as shown in Fig. 1(ii). Each link carries a control wavelength w_0 and n data wavelengths $(w_1, w_2, .., w_n)$. The basic architectural component is a 1x2 variable optical splitter (VOS). Other components are combiners, multiplexers (MUX), demultiplexers (DEMUX), receivers (Rx), transmitters (Tx), and a control processing unit. In Fig. 1(i), the components VOSs, combiners, receivers, and transmitters are shown for the data wavelength w_1. VOSs are cascaded and linked to combiners and receivers as shown in the figure. Additional VOSs and combiners can be cascaded and linked in the similar manner to accommodate more links. For F number of fiber links and N data wavelengths, a total of NF^2 VOSs and NF combiners (each is of type $(F+1)X1$) are required.

We use the 1x2 VOS component presented in [14] [15] [13] in our switch. The self-latching VOS is based on magneto-optical technology. The VOS is designed using mainly a variable faraday rotator and a walk-off crystal. In the VOS, input optical power can be distributed (or split) on the two output ports with various ratios (states) such as $(0\% - 100\%)$, $(50\% - 50\%)$, and $(100\% - 0\%)$ as shown in Fig. 1(iii). The component requires an electric pulse to switch states (i.e. increase/decrease the power on a port). By applying the electric pulse appropriately the various states can be achieved. It takes 0.25ms time to switch between $(0\% - 100\%)$ and $(50\% - 50\%)$ states. We assume the same time period to switch between $(50\% - 50\%)$, and $(100\% - 0\%)$ because of near symmetrical power splitting pattern seen in [15]. We denote the 0.5ms configuration time required to change the split power

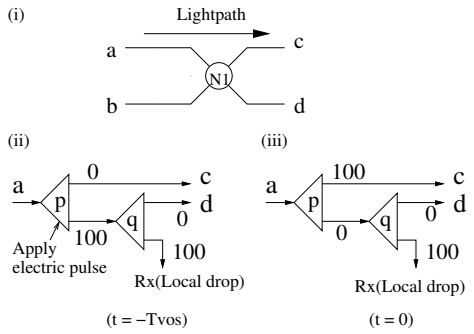


Fig. 2. Switch configuration

on an output port from 0% to 100% (i.e., from $(0\% - 100\%)$ to $(100\% - 0\%)$) as $Tvos$. The states $(0\% - 100\%)$ and $(100\% - 0\%)$ can be used to direct the full power towards one output port only (like the traditional optical switches). Further, the state $(50\% - 50\%)$ can be used to split power. Particularly, unlike the traditional switches, the power can be split on a desired sub-set of output ports on need basis by appropriately changing the state of VOSs in the switch. We use this feature in our protection approach, which is illustrated in Section III. The low-cost magneto-optic component available in [16] can also be used in our architecture.

Average insertion loss (IL) of the VOS is 0.6dB (for $(0\% - 100\%)$ and $(100\% - 0\%)$) and 4dB (for $(50\% - 50\%)$) and polarization-dependent loss (PDL) is less than 0.1dB. The VOS energy consumption is very low ($\sim 120\mu J$) [14]. For a typical nodal degree such as two and three, the insertion loss at core components (VOSs and combiners) is in the range of 6.6dB to 7.2dB and 6.6dB to 7.8dB respectively (VOSs: 1.2dB and 1.8dB when traversing up to 2 and 3 VOSs with $(0\% - 100\%)$ and $(100\% - 0\%)$ states, and combiners: 6dB when two cascaded combiners (each of 3dB type) are traversed with these nodal degrees in a 4X1 type). When the VOS is used with $(50\% - 50\%)$ state (used when failure recovery only), slightly more power loss occurs. Therefore, the architecture is suitable for networks in small geographical areas because of the power-loss. Otherwise, amplifiers are required to compensate the power-loss.

A. Switch configuration

An optical connection/lightpath can be set up by configuring intermediate nodes along the lightpath. Generally, control messages are sent (on the control wavelength w_0) using a two-way reservation approach for establishing the lightpath (on a data wavelength, say w_1). The control message is processed electronically at the control processing unit at intermediate nodes. The control message carries the details about the connection which are used to configure VOSs at intermediate nodes. Below, we illustrate how VOSs are configured at a node which connects an input port to an output port for establishing the lightpath. We consider a node with two input and two output links as shown in Fig. 1(ii) for illustration.

We consider that a lightpath is set up which traverse from link a to link c on the wavelength w_1 as shown in Fig. 2(i). It is considered that, at $t = -Tvos$, the control message has been processed and the switch configuration is initiated. The default status of VOSs in our switch (at $t = -Tvos$) is shown in Fig. 2(ii). VOS-p and VOS-q shown are the two VOSs connected with link a in our switch architecture shown in Fig. 1(i). We do not show the other VOSs connected with link b as no configuration is done in these VOSs. At $t = -Tvos$, the default power splitting status of both VOS-p and VOS-q is $(0\% - 100\%)$ with 0% power directed towards links c and d. That is, paths within the switch from a to c and a to d are shut initially. Once the control message has been processed, the node identifies the output port of the connection and selects the VOS which is connected to that port (i.e. VOS-p). An electrical pulse is applied to the selected VOS, i.e. VOS-p, at $t = -Tvos$ as shown in Fig. 2(ii). It changes the power splitting state of VOS-p to $(100\% - 0\%)$ at $t = 0$ (i.e. it requires $Tvos$ time to change the state) with 100% power directed towards links c. This is shown in Fig. 2(iii). That is, power directed towards link c increases from 0% (at $t = -Tvos$) to 100% (at $t = 0$). Therefore, at $t = 0$, the path $a - c$ is connected/opened. When optical signals arrive on the lightpath they are switched with full input power directed towards link c. Paths $a - d$ remains shut. Note that, at the receiver node (egress), the optical signals can be received at the default state as full power is directed towards the local receiver (Rx).

III. Pre-configured backup protection with limited sharing (*PBPLS*)

The traditional optical layer dedicated protection has short recovery time because of pre-configured backup paths. Achieving short recovery time by pre-configured backup paths and at the same time employing backup sharing are not done. This is because of the limitations in the traditional OXCs. Consider that a switch configuration is done to connect an input port to an output port within a widely used OXC such as a MEMS optical switch. While maintaining this connection, another configuration to connect (1) the same input port to a different output port, or (2) a different input port to the same output port is not done. This is because, this later configuration disrupts the existing connection. The configuration is, therefore, done only after the existing connection is over or released. This constraint does not allow setting up two backup lightpaths which are pre-configured and at the same time they share one or more wavelength links.

As explained in Section II, the switch architecture considered in this paper has increased flexibility of how optical power received on an input port can be directed or split on need basis. This flexibility can be used to overcome the above constraint. Power splitting allows connecting an input port to two (or more) output ports within the switch. In addition to this, the components are cascaded in the architecture such that they allow configurations which connect two (or more) input ports to the same output port. We illustrate how these features are used in our protection approach below.

 ISBN:978-1-61208-113-7

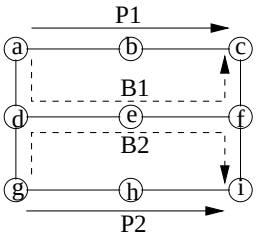


Fig. 3. Shared protection

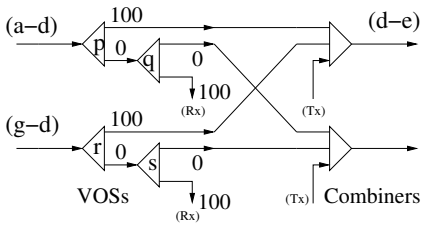


Fig. 4. Switch configuration at node-d

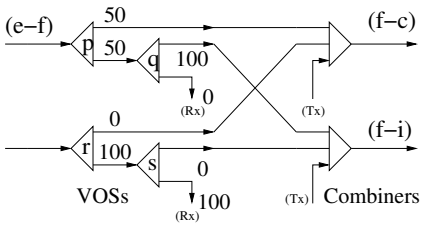


Fig. 5. Switch configuration at node-f

The protection approach allows provisioning pre-configured backup paths with limited backup resource sharing. Such backup sharing is possible in single component failure scenarios. We assume single link failures which are the predominant type of component failures. The proposed approach is illustrated in Fig. 3. It shows two primary lightpaths (P1 and P2) and their backup lightpaths (B1 and B2). Each of the backup lightpaths is link-disjoint with its primary lightpath. Further, primary lightpaths P1 and P2 are also link-disjoint as shown. Links $d-e$ and $e-f$ are shared among the backup paths. When using the proposed switch, configurations can be done at node *d* such that the link $d-e$ can be opened for both the links $a-d$ and $g-d$ for transmission at the same time. This is shown in Fig. 4 (The same switch shown in Fig. 1(i) is used for this illustration. Only the VOSs and combiners for wavelength w_1 are shown in Fig. 4. The additional output link which is not labeled in the figure is not used). In this configuration, VOS-p and VOS-r are configured such that their power splitting state becomes $(100\%-0\%)$ with 100% power directed towards link $d-e$. The configurations are done by applying electric pulses as explained in Section II-A. Further, at node *f*, power from $e-f$ can be split on $f-c$ and $f-i$. This configuration is shown in Fig. 5. In this configuration, VOS-p is configured to the splitting state $(50\%-50\%)$ and VOS-q is configured to the splitting state $(100\%-0\%)$ (100% directed towards link $f-i$). As a result of these configurations, the power from $e-f$ is split on $f-c$ and $f-i$. While sharing backup links $d-e$ and $e-f$, these configurations allow transmission over a backup path without needing further configuration. No power splitting occurs at VOSs at nodes *d* and *e* (at node *e*, similar configuration illustrated in Section II-A is done). Note that, the above configurations are done at the time when the primary connections are established.

In case of failure on P1, traffic can be immediately rerouted through B1 since it is pre-configured. The traffic will be switched from $a-d$ to $d-e$ because of the switch configuration illustrated above. Further, the traffic will be switched from $e-f$ to $f-c$ because of the power splitting configuration. Hence, it provides short recovery time which is equivalent to the case of dedicated protection. When rerouting the traffic, a copy of traffic is routed on the link $f-i$ also due to power splitting. Similar rerouting can be done when failure occurs on P2. Note that, in Fig. 5, power from $e-f$ is split towards the desired output links $f-c$ and $f-i$ only, and power wastage can be reduced by not splitting on unwanted ports (if any).

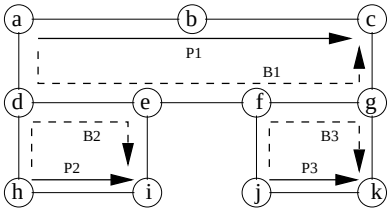


Fig. 6. A backup sharing scenario

A. Limited backup sharing

Backup sharing while provisioning pre-configured backup paths is limited because of power loss due to power splitting. We mainly consider splitting loss at VOSs. Repeated or cascading of power splitting may occur when backup links are shared by many backup lightpaths. This is illustrated in Fig 6. Three primary lightpaths (P1, P2, and P3) are protected by their pre-configured backup paths (B1, B2, and B3). Links $d-e$ and $f-g$ are shared by backup lightpaths B1 & B2, and B1 & B3 respectively. In case of failure on P1, traffic is rerouted via B1. In this case, power splitting occurs at nodes e and g (at VOSs). To reduce the power loss due to such repeated power splitting, we adopt the approach of limiting the number of power splitting at intermediated nodes (at VOSs) on a backup lightpath. Limiting the number of power splitting limits the degree of backup sharing. We denote the maximum number of power splitting at intermediate nodes on a backup lightpath as *K*. For instance, the three backup lightpaths can be provisioned in Fig 6 when $K=2$. However, if $K=1$, only primary lightpaths P1 & P2 can be set up with backup paths B1 & B2 respectively. Lightpath P3 has to be rejected since B3 would, otherwise, cause additional power splitting at node *g*. Similarly, once P1 and P2 have been admitted with their backup lightpaths, consider admitting a future request with its primary and its backup lightpaths (say, P4 and B4 (not shown)). Assume that B4 shares the same backup link $d-e$ and additional power splitting occurs at node e to a link (say $e-l$) in addition to the links $e-f$ and $e-i$ (the link $e-l$ is not shown). In this scenario, with $K=1$, this new request is rejected since additional power splitting occurs.

B. Protection with fixed splitters vs. VOS

Our proposed protection approach can also be implemented with traditional (fixed) splitters and shutters instead of using VOSs. (Similar splitter-shutter type switches are broadcast and

select based switches [8] considered for optical burst/packet switching networks) A major drawback with fixed splitters based switches is their high power loss. This is because power is always split towards all the output ports. Therefore, only a small portion of power is used to transmit data and the remaining power is wasted. This small power may not be enough for transmission over long distance. In addition to this, additional shutters are required. With VOSs, even with a large number of ports, optical signals are switched with 100% power directed at VOSs towards the output link during normal working conditions. In case of failure-recovery using shared backup lightpaths, optical power is split towards necessary output ports only. Therefore, power-wastage is significantly reduced. In addition to this, additional shutters are not required when VOSs are used. Because of these reasons, we use VOSs instead of traditional splitters.

IV. Performance Study

We evaluate the performance of the proposed protection approach (PBPLS) on the 14 node and 21 bi-directional link NSFNET topology. We consider 16 wavelengths per fiber. We consider sub-lambda connection requests (or LSPs) which require optical layer protection. A sub-lambda connection can traverse a number of lambda connections or lightpaths. In the optical layer protection, each of the lightpaths traversed is protected by a backup lightpath. Traffic requests arrive dynamically. Request arrivals follow Poisson distribution and holding time of a request follows exponential distribution with unit mean. We assume wavelength capacity to be 10 units. Bandwidth requests for traffic are uniformly distributed in the range of (4-10). Each request's source node and destination node are selected based on uniform distribution. We use a shortest path selection algorithm (Dijkstra's algorithm) with the objective of minimizing the total number of physical hops to route the requests. Each experiment is carried out with a large number of request arrivals on the order of 10^5.

We investigate whether significant performance improvement is seen when limiting the number of power splitting at intermediate nodes (K) to small values ($K = 1$, $K = 2$, and $K = 3$). First, we consider that all the traffic requests require short recovery time and they are protected with pre-configured backup paths using our proposed protection approach, PBPLS. We compare the performance with the traditional dedicated protection since it also provides pre-configured backup paths (recovery time in PBPLS is equivalent to the case of the traditional dedicated protection). In addition to this, we also study the performance with two classes of traffic when only a portion of requests require short recovery time (class-1) while the rest can tolerate slightly longer recovery time (class-2). For class-1, pre-configured backup paths are provided using PBPLS (and compared with the traditional dedicated protection). For class-2, non pre-configured backup paths are given using the traditional optical layer shared protection approach. In this study, two traffic arrival distributions are considered. The traffic arrival follows the distribution, class-1 : class-2 = (1) 50% : 50%, and (2) 25% : 75%. In this study, we

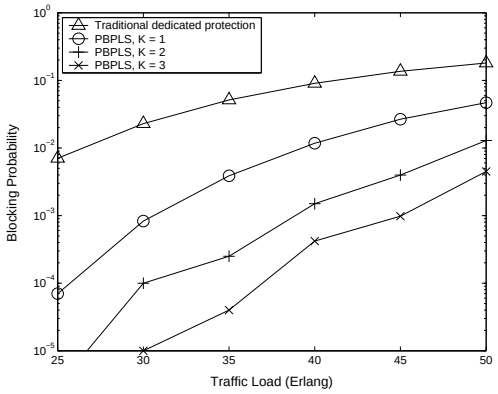


Fig. 7. Performance for the traditional and proposed protection approaches

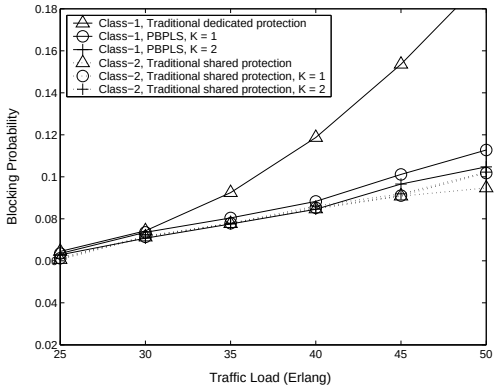


Fig. 8. Performance of class-1 (50%) and class-2 (50%) requests

also observe whether the performance improvement for class-1 traffic due to limited backup sharing penalizes class-2 traffic. When both the PBPLS and the traditional optical layer shared protection are provided, we consider that backup resources associated to these two protections are separated (i.e., pre-configured backup paths and traditionally shared backup paths (not pre-configured) do not share the same resources) in order to reduce the complexity.

The blocking performance for the proposed PBPLS approach with different values of the number of power splitting ($K = 1$, $K = 2$, and $K = 3$) and the traditional dedicated protection approach are shown in Fig. 7. In this study, a single class of traffic is considered and all the requests are admitted using the same protection method. It can be seen that, significant reduction in blocking is achieved in PBPLS with $K = 1$ (more than 74% reduction in blocking when compared to the traditional approach). This is because, significantly a large number of requests can find backup resources as resources can be shared though it is limited in our approach. Further reduction in blocking is observed with increasing number of power splitting (with $K = 2$ and $K = 3$, additional 18% and 23% blocking reduction is seen at high loads).

Figure 8 shows the blocking performance when 50% of requests (class-1) are protected by pre-configured backup paths (PBPLS is used. The traditional dedicated protection is used

 ISBN:978-1-61208-113-7

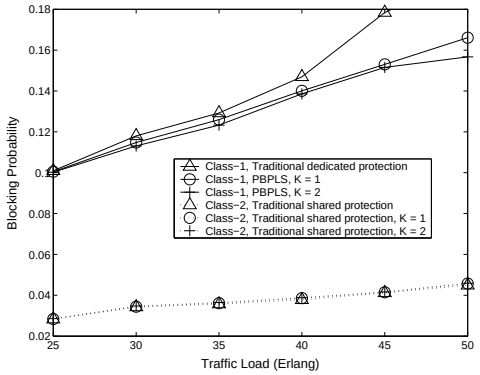


Fig. 9. Performance of class-1 (25%) and class-2 (75%) requests

for comparison), and 50% of requests (class-2) are protected by non pre-configured backup paths (the traditional optical layer shared protection approach is used). Two values for the number of power splitting ($K = 1$, and $K = 2$) are investigated. In the figure, we denote the performance of class-2 traffic when class-1 traffic is admitted using PBPLS with $K = p$ as 'Class-2, Traditional shared protection, $K = p$'. We denote the performance of class-2 traffic when class-1 traffic is admitted using the traditional dedicated protection as 'Class-2, Traditional shared protection'. For class-1 requests, up to about 42% reduction in blocking is seen for PBPLS with $K = 1$, when compared to the traditional dedicated protection approach. With $K = 2$, additional blocking reduction of up to 4% only is seen. The performance of class-2 requests is shown in dotted lines. The impact on the performance of class-2 requests as a result of using our proposed protection for class-1 requests is seen at very high loads only (above 45 Erlang). For class-2 requests, about 7% additional blocking is seen at 50 Erlang when $K = 1$ in our protection approach.

Figure 9 shows the blocking performance when 25% of requests (class-1) are protected by pre-configured backup paths and 75% of requests (class-2) are protected by non pre-configured backup paths. Overall, more blocking is seen for class-1 requests. This is because, more resources are occupied by frequently arriving class-2 requests. Class-1 requests do not arrive frequently with the given small percentage of traffic arrival. Therefore, they may not find enough available resources and they are blocked. Even with the small percentage of traffic arrival, considerable blocking reduction of up to 15% is seen for class-1 requests with the proposed protection with $K = 1$. The impact on the performance of class-2 requests as a result of improved performance for class-1 requests is not significant (only about 2% additional blocking is seen for class-2 requests)

V. Conclusions

In this paper, we proposed an optical protection approach with pre-configured backup paths, which allows limited backup resource sharing. We investigated the performance in a single class (all the requests were provisioned with pre-configured backup paths) and two-class (class-1 and class-2 requests were provisioned using pre-configured and non pre-configured backup paths respectively) traffic scenarios. We demonstrated that even with the small number of power splitting (K), significant performance improvement is seen. In the single class scenario, our proposed approach with *K=1* showed more than 74% reduction in blocking when compared to the traditional dedication protection approach. In the two-class scenario, it showed up to 42% and 15% reduction in blocking for class-1 traffic with 50% and 25% traffic distributions respectively.

Acknowledgment

This research work was supported by NUS ARF research grant R-263-000-530-112.

References

[1] V. Anand and C. Qiao, "Dynamic Establishment of Protection Paths in WDM Networks. Part-I," in Proceedings of *IEEE ICCCN*, pp. 198-204, Oct. 2000.

[2] S. Ramamurthy, L. Sahasrabuddhe, and B. Mukherjee, "Survivable WDM Mesh Networks," *IEEE Journal of Lightwave Technology*, vol. 21, no. 4, pp. 870-882, Apr. 2003.

[3] C. Ou et al., "Traffic Grooming for Survivable WDM Networks - Dedicated Protection," *OSA Journal of Optical Networking*, vol. 3, pp. 50-74, Jan. 2004.

[4] A. Mykkeltveit, and B. E. Helvik, "Comparison of Schemes for Provision of Differentiated Availability-Guaranteed Services Using Dedicated Protection," in Proceedings of *IEEE International Conference on Networking, ICN*, pp. 78 - 86, 2008.

[5] M. Liu, M. Tornatore, and B. Mukherjee, "New and improved strategies for optical protection in mixed-line-rate WDM networks," in proceedings of *OFC 2010*, 2010.

[6] R. Krishanthmohan, G. Mohan, and K. C. Chua, "A Flexible Optical Switch Architecture for Efficient Transmission of Optical Bursts," under review, *Computer Networks Journal*.

[7] Ai-Qun Liu, *Photonic MEMS devices: Design, Fabrication and Control*, CRC Press, 2009.

[8] C. Guillemot et al., "Transparent optical packet switching: the European ACTS KEOPS project approach," *Journal of Lightwave Technology*, vol. 16, no. 12, pp. 2117 - 2134, Dec. 1998.

[9] M. C. Yuang et al., "A QoS optical packet switching system: architectural design and experimental demonstration," vol. 48, no. 5, pp. 66 - 75, May 2010.

[10] R. Krishanthmohan, G. Mohan, and Z. Luying, "Differentiated Survivability with Improved Fairness in IP/MPLS-over-WDM Optical Networks," *Computer Networks Journal*, vol. 53, no. 5, pp. 634-649, Apr. 2009.

[11] J. K. Rhee, I. Tomkos, and M. J. Li, "A broadcast-and-select OADM optical network with dedicated optical-channel protection," *IEEE Journal of Lightwave Technology*, vol. 21, no. 1, pp. 25 - 31. Jan. 2003.

[12] L. Long, and A. E. Kamal, "Tree-Based Protection of Multicast Services in WDM Mesh Networks," in Proceedings of *IEEE GLOBECOM 2009*, 2009.

[13] S. Wada, S. Abe, Y. Ota, B. Reichman, T. Amura, and C.A. Daza, "Variable gain equalizer using magneto-optics," in proceedings of *OFC 2002*, vol. 70, pp. 324 - 326, 2002.

[14] H. Ramanitra, P. Chanclou, J. Etrillard, Y. Anma, H. Nakada, and H. Ono, "Optical access network using a self-latching variable splitter remotely powered through an optical fiber link," *Optical Engineering*, vol. 46, Apr. 2007.

[15] H. Ramanitra, P. Chanclou, Z. Belfqih, M. Moignard, H. L. Bras, and D. Schumacher, "Scalable and multi-service passive optical access infrastructure using variable optical splitters," in proceedings of *OFC 2006*.

[16] Agiltron Inc. [Online]. Available: http://www.agiltron.com/PDFs/CL1x2 multicasting switch.pdf

 ISBN:978-1-61208-113-7

DNS Security Control Measures: A heuristic-based Approach to Identify Real-time incidents

Joao Afonso
Foundation for National Scientific Computing
Lisbon, Portugal
e-mail: joao.afonso@fccn.pt

Pedro Veiga
Department of Informatics
University of Lisbon
Lisbon, Portugal
e-mail: pedro.veiga@di.fc.ul.pt

***Abstract*—There is no doubt that one of the most critical components of the Internet is the DNS – Domain Name System. In this paper, we propose a solution to strengthen the security of DNS servers, namely those associated with Top Level Domains (TLD), by using a system that identifies patterns of potentially harmful traffic and isolates it. The proposed solution has been developed and tested at FCCN, the TLD manager for the .PT domain. The system consists of network sensors that monitor the network in real-time and can dynamically detect, prevent, or limit the scope of the attempted intrusions or other types of attacks to the DNS service, thus improving it's global availability.**

***Keywords*—DNS ; security; intrusion detection system; real-time;monitoring.**

I. INTRODUCTION

The DNS protocol is the basis of a critical Internet application used for the reliable and trustworthy operation of the Internet. DNS servers assume a central role in the normal functioning of the Internet by resolving domain names into network addresses for IP networks. Any disturbance to their normal operation can have a dramatic impact on the service they provide and on the global Internet. Although based on a small set of basic rules, stored in files, and distributed hierarchically, the DNS service has evolved into a very complex system and critical system [1].

According to recent studies [2], there are nearly 11.7 million public DNS servers on the Internet. It is estimated that nearly 52% of them, due to improper configuration, allow arbitrary queries (thus allowing denial of service attacks or "poisoning" of the cache). About 31.1% of the servers also allow for the transfer of their DNS zones.

There are still nearly 33% of situations where the authoritative nameservers of an area are on the same network, which facilitates the attacks of the type of Denial of Service (DOS), a frequent attack to the DNS. Furthermore, the type of attacks targeting the DNS is becoming more sophisticated, making them more difficult to detect and control on time. Examples are the attacks by Fast Flux (ability to quickly move the DNS information about the domain to delay or evade detection) and its recent evolution to Double Flux.

One of these attacks, is the conficker [3] worm, first appeared on October 2008, but also known as Code Red, Blaster, Sasser and SQL Slammer. Every type of computer, using a Microsoft Operating System can potentially be infected. Attempts to estimate the populations of conficker have lead to many different figures but all these estimates exceed millions of personal computers. Conficker made use of domain names instead of IP address in order to make its attack networks resilient against detection and takedown.

The ICANN - Internet Corporation for Assigned Names and Numbers, created a list containing the domains that could be used in each TLD in such attacks to simplify the work of identifying attacked domains.

A central aspect of the security system that we propose and have implemented is the ability to collect statistically useful data about network traffic for a DNS resolver and use it to identify classes of harmful traffic to the normal operation of the DNS infrastructure. In addition to collecting data the system can take protective actions by detecting trends and patterns in the traffic data that might suggest a new type of attack or simply to record important parameters to help improve the performance of the overall DNS system.

The fact that the DNS is based on an autonomous database, distributed by hierarchy, means that whatever solution we use to monitor, it must respect this topology. In this paper we propose a distributed system using a network of sensors, which operate in conjunction with the DNS servers of one or more TLDs, monitoring in real-time the data that passes through them and taking actions when considered adequate.

The ability to perform real-time analysis is crucial in the DNS area since it may be necessary to immediately act in case of abuse or attack, by blocking a particular access and notifying other cooperating sensors on the origin of the problem, since several types of attacks may be directed to other DNS components. The use of a Firewall solution whose triggering rules are dynamically generated by the network sensors is a fundamental component of the system, to filter attacking systems in an efficient way and resuming to the initial situation when the reason to filter different traffic patterns has ceased to exist. With this approach we aim to guarantee an autonomous functioning of the platform without the need of human intervention.

 ISBN:978-1-61208-113-7

The use of network alarms can also help in monitoring the correct functioning of the whole solution. Special care has been taken to minimize the detection of false positives or also false negatives.

The remaining of the paper is structured as follows: Section II provides background information regarding related work. Section III introduces out proposed methodology. In section IV, we describe the solution. Section V presents a case study for validation of the proposal. In Section VI, the results gathered in the case study are analyzed. Finally, Section VII presents some conclusions and directions for further work.

II. RELATED WORK

One of the first studies that can be observed in this area has the authorship of Guenter and Kolar, with a tool called sqldjbdns [4]. Their proposal uses a modified version of the traditional BIND [5] working together with a Structured Query Language (SQL) version inside a Relational database management system (RDBMS). For DNS clients, this solution is transparent and there is no difference from classic BIND.

Zdrnja presented a system for Security Monitoring of DNS traffic [6], using network sensors without interfering with the DNS servers to be monitored. This is a transparent solution that does not compromise the high availability needed for the DNS service.

Vixie proposed a DNS traffic capture utility called, DNSCap [7]. This tool is able to produce binary data using pcap format, either on standard output or in successive dump files. The application is similar to tcpdump [8] – command line tool for monitoring network traffic, and has finer grained packet recognition tailored for DNS transactions and protocol options, allowing for instance to see the full DNS message when tcpdump only shows a one-line summary.

Another tool available is DSC - DNS Statistics Collector [9]. DSC is an application for collecting and analyzing statistics from busy DNS servers. Major features include the ability to parse, summarize and search inside DNS queries detail. All data is stored in an SQL database. This tool, can work inside a DNS server or in another server that "captures" bi-directional traffic for a DNS node.

Kristoff also proposed an automated incident response system using BIND query logs [10]. This particular system, besides the common statistical analysis, also provides information regarding the kind of consultations operated. All information is available through the Web based portal. Each security incident can result in port deactivation.

III. METHODOLOGY

A. Architecture

The architecture of the system that we have develop aims to improve the security, performance and efficiency of the DNS protocol, removing all unwanted traffic and reinforce the resilience of a Top Level Domain. We propose an architecture comprising an integrated protection of multiple DNS servers, working together with several network sensors that apply live rules to a dedicated firewall, acting as a traffic shaping element.

Sensors located carefully located in the network monitor all the traffic going to the DNS infrastructure, identify potentially harmful traffic using a algorithm that we have developed and tested and use this information to isolate traffic that has been identified has security threats.

Several networks sensor monitor different parts of the infrastructure and exchange information related to security attacks. In this way, as shown in Fig. 1, it should also be possible to exchange critical security information between the sensors. In addition to an increase in performance, this operation should prevent an attack on a server from a source, identified by another sensor as malicious. This scenario is relevant since some kinds of attacks are directed to several components of the DNS infrastructure.

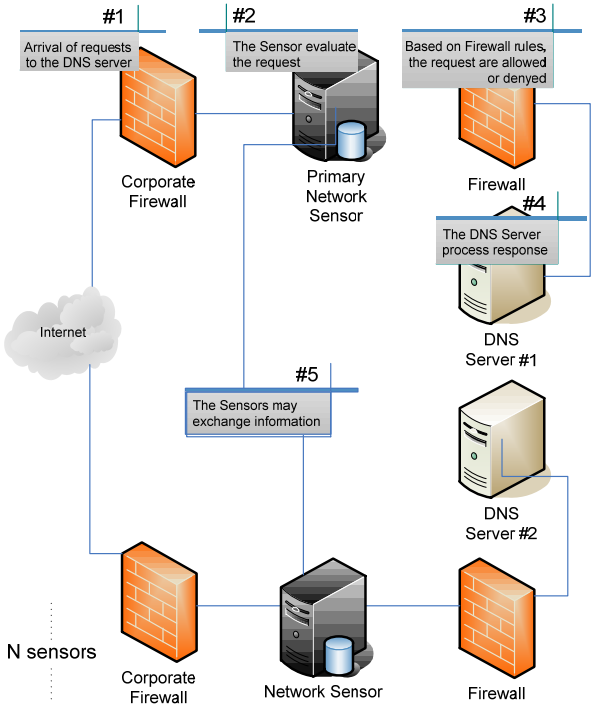


Figure1. Diagram of the desired solution

B. Heuristic

One of the crucial parts of our work is the algorithm to identify traffic harmful to the DNS. In order to implement the stated hypothesis in the architecture and keep the DNS protocol as efficient as possible, it is necessary to apply a heuristic, which in real time, evaluates all the information collected from different sources and applies convenient weights to each component and act accordingly.

The components that we have chosen to have impact in the security incidents of DNS are: the number of occurrences, analysis of type of queries been made, the amount of time between occurrences, the number of probes affected and information reported from intrusion detection systems.

 ISBN:978-1-61208-113-7

Our system uses the following formula to evaluate a parameter that measures the likelihood of the occurrence of a security incident:

$$f(x) = O \cdot 0{,}2 + C \cdot 0{,}2 + G \cdot 0{,}15 + N \cdot 0{,}25 + I \cdot 0{,}20$$

Are factors considered in applying this formula:

- Occurrences (O) - Represents the number of times (instances) that have given source was blocked, so that the distributed then depicted in Table I.

TABLE I – CONTRIBUTION OF THE NUMBER OF OCCURRENCES OF A SOURCE IN MALICIOUS HEURISTIC

Occurrences	*Weight*
1	25%
2	50%
3	75%
4 or more	100%

- Analysis (C) - Real-time evaluation of the deviation of the values recorded in relation to the average observed statistics, based on the criteria and weights identified below in Table II.

TABLE II – CONTRIBUTION OF EVENTS TYPIFIED A POTENTIALLY MALICIOUS SOURCE GIVEN IN HEURISTIC

Event	*Weight*
Entire zone transfer attempt (AXFR)	100%
Partial transfer zone attempt (IXFR)	50%
Incorrect query volume, 50 to 75% on average per source	75%
Incorrect query volume exceeding 75%	100%
Query volume, up 50%, the average number of access by origin	50%

Note that the estimates apply the moving average, for the determination of reference values, given the ongoing development of data collected.

- Time between occurrences (G) - time since last occurrence of a given source, distributed with the weights associated to the times below are obeisant.

TABLE III – WEIGHT OF DIFFERENT TIME BETWEEN EACH OCCURRENCE

Time	*Weight*
Less than 1 Minute	100%
Less than 1 Hour	75%
Less than 1 Day	50%
Less than 1 Week	25%

- Incidence (N) - Number of probes that report blocks in the same source.
  For the calculation, we observed expression:

$$\frac{1}{\#Total\_Sensors - \#Sensors\_Attacked}$$

- Intrusion Detection Systems (I) - We considered the use of the Snort platform, being free to use, and gather a large number of notarized signatures of security incidents relating to the DNS service.

TABLE IV – INTERCONNECTION WITH TEMPORAL DATA GATHERED FROM INTRUSION DETECTION SYSTEMS

Metric: Common Vulnerability Scoring System (CVSS)	*Weight*
Low level	34%
Middle level	67%
High level	100%

For the activation of a rule in Firewall occurs will require:

1. The formula shown above take values equal to or greater than 0.25;
2. The combination of two or more criteria of the formula.
   Exception: when receiving information from all the other sensors, in which case a single criteria is sufficient;
3. It respected the existing white list in the repository, allowing considered privileged sources that are not blocked.
   In this way we avoid compromising the Internet service, considering the key role played by DNS, the White List protects key addresses from being blocked in case of false positives events.
   This list is created from a record of trusted sources, allowing all addresses listed here to be protected from being added to the Firewall rules.
   One example is the list of internal addresses, and the DNS servers of ISPs.

Instead, for the removal of a rule in the firewall will need to occur simultaneously on the following assumptions:

1. Exceeded the quarantine period, based on the parameters in use;
2. The expression of activation (heuristic) does not (still) check the referenced source.

 ISBN:978-1-61208-113-7

IV. PROPOSED SOLUTION

A. Diagram

As shown in Fig. 2, this solution is based on a network of sensor engines that analyze all traffic flowing into the DNS server in the form of valid or invalid queries, process the information received from other probes and issue restrictions for specific network addresses. In case an abnormal behavior is detected or there is suspicious behavior from a certain network address, it will be blocked in the firewall and the other probes notified so they can act accordingly. The system can also calculate the response time for each operation to evaluate the performance of the server.

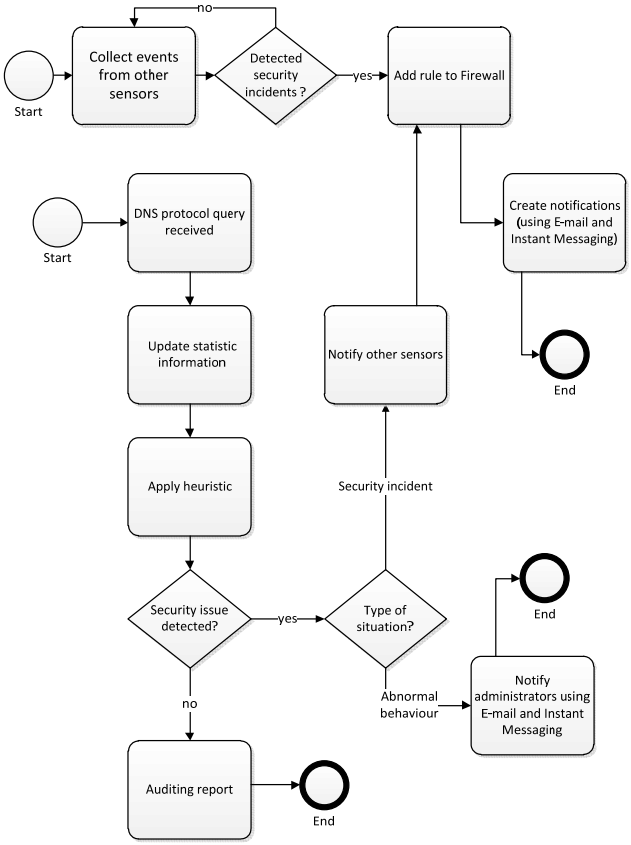


Figure 2. Block Diagram of proposed solution

For each rule inserted in the sensor firewall, there will be a period of quarantine and, at the end of this time, the sensor will evaluate the behavior of that source, to evaluate the needed to remove the rule, as shown in Fig. 3.

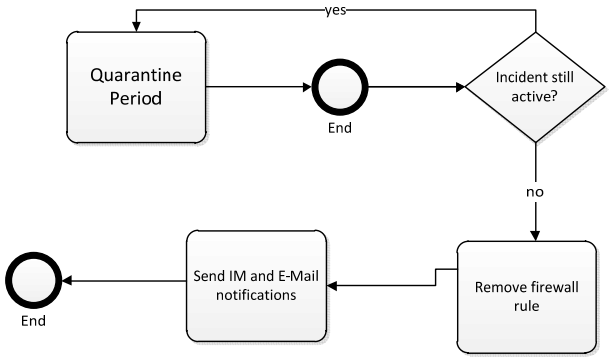


Figure 3. Quarantine procedure over the Firewall

B. Network data flow

According to our design, all data that flows through the probe heading for the DNS server is treated according to a standard set of global firewall rules, followed by specific local rules regarding to the addresses that are being blocked in real time. The queries are then delivered to the parser to be analyzed and stored in the RDBMS. At the top is the system of alarms and the Web portal (Fig. 4).

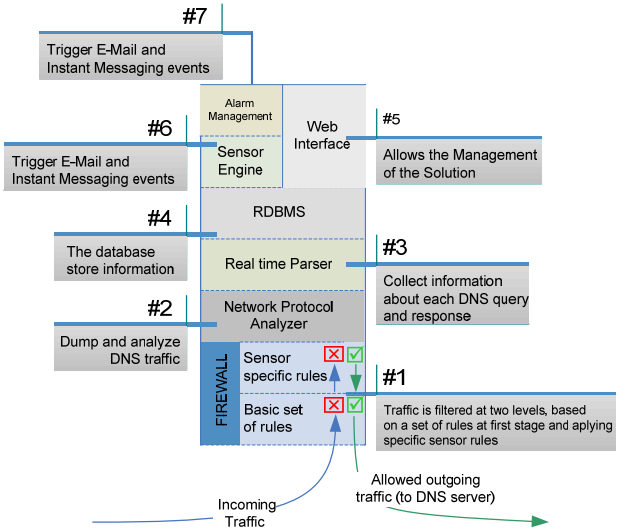


Figure 4. Network data flow

All information collected is stored in a database implemented in MySQL [11]. Taking into consideration the need to optimize the performance of the queries and to reduce the volume of information stored, the data is divided into a number of different tables.

The conversion of the IP address of source and destination (DNS server) into an integer format, has allowed for much more efficient data storage, and significant improvements in the overall performance of the solution.

 ISBN:978-1-61208-113-7

The information regarding all queries made, is stored daily into a log, and kept available during the next 30 days.

Two tables containing the set of rules that are dynamically applied – add or removed, based on situations that have been triggered - control the correct operation of the firewall. For auditing purposes every action is registered.

The information required for auditing and statistical tasks never expires.

C. *Statistical analysis and performance evaluation*

The statistical information collected and stored in the database has a significant amount of detail. It is possible, for example, to calculate, for each sensor, the evolution of queries per unit of time (hour, day, etc) badly formatted requests, DNS queries of rare types and determine the sources that produce the larger number of consultations. It is also possible to see the standard deviation of a given measure so we can relate it to that is seen with the other hits [14].

The performance of the DNS protocol responses is permanently measured, regarding the response time per request. Data is constantly registered and an alarm is raised in case normal response times are exceeded.

V. CASE STUDY

Our proposal have been under development since September 2006 at FCCN – who has the responsibility to manage, register and maintain the domains under the .PT TLD.

At present time, there are two sensors running attached to the DNS servers (one at the primary DNS and another working together with a secondary DNS server).

The network analyzer is tshark [15], and the firewall used is IPFilter [12]. The real time parser was programmed in Java, collecting the information received from the tshark. The Web server is running Apache with PHP.

Regarding the Xmpp server [13] we choose the Jive messenger platform.

All modules are integrated together.

The entire sensor solution, as described above, as well as the web platform we developed went on-line on the 1st of January 2007, and the data from the various agents was collected from the 10th of May 2008 till now.

VI. RESULTS

We present here the results of the last 12 months of data collection (between 1st of May 2009 and 31st May 2010). The Average number of requests to the primary DNS server is up to 14,459,356 per day (167 per sec.).

The performance of the data analysis program is above 1240 requests processed per sec. (filtered, validated and inserted in the database).

Using the data collected by the sensors, during this time period, we were able to:

- Collect useful statistical information. E.g., daily statistics by type of DNS protocol registers accessed (Fig. 5).

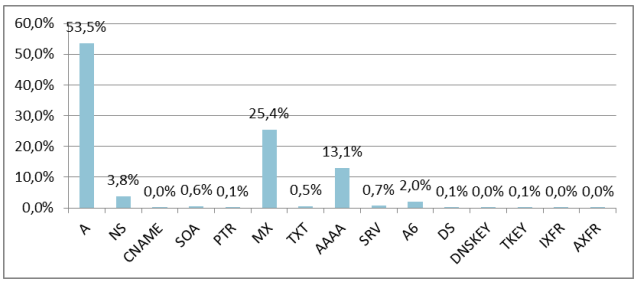


Figure 5. Statistical analysis by type of records accessed

- Detect examples of abnormal use (that are not security incidents). For example we were able to detect that a given IP was using the primary .PT DNS server as location resolver.
  The number of queries made was excessive when compared with the average value per source, reaching values close to some Internet Service Providers that operate under the .PT domain.
- Detect situations of abuse, including denial of service attacks, with the execution of massive queries. In last 12 months of analysis there are 17 DOS attacks triggered.
  They were instantly blocked, and addresses placed in quarantine (Table V).

TABLE V. EXAMPLES WHEN THE SENSOR DETECTED SITUATIONS THAT REQUIRED THE FIREWALL RULES TO CHANGE.

Source Address	*Date / Time*	*Operation*	*Sensor*
xx.xx.200.35	2010-04-15 02:05:04	Add rule	xx.xx.44.62
xx.xx.17.212	2010-04-15 03:15:02	Remove rule	xx.xx.44.63
xx.xx.117.51	2010-04-15 03:47:24	Add rule	xx.xx.44.63
xx.xx.94.139	2010-04-15 04:27:19	Add rule	xx.xx.44.62
xx.xx.13.231	2010-04-15 07:35:58	Remove rule	xx.xx.44.62

- Improve DNS protocol performance repairing situations of inefficient parameterization of the DNS server.
  On the DNS server side, considering the capacity of the probe to determine the processing time for each consultation, it is possible to detect cases of excessive delay, which was later confirmed to coincide with of moments of zone update (Fig. 6).

 ISBN:978-1-61208-113-7

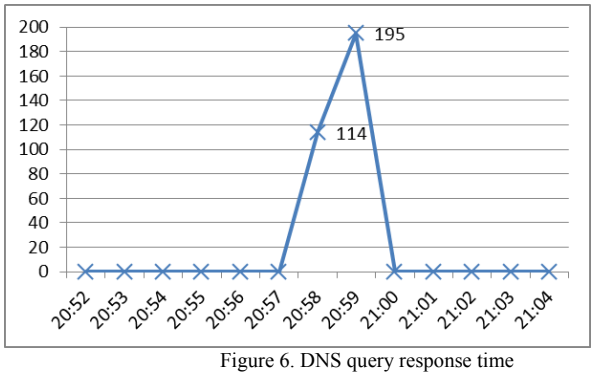


Figure 6. DNS query response time

Considering the daily progress of DNS queries, before and after applying shaping heuristic to the protocol we obtain an improvement between values of 5.3% (minimum) and 19.4% (maximum), as witnessed in Fig. 7.

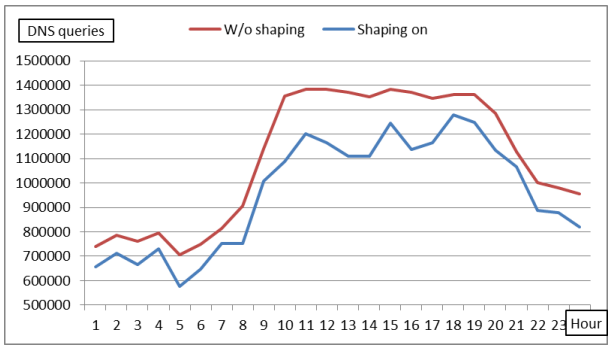


Figure 7. Improvement of DNS performance protocol

Fig. 8 shows the recurrence of same IP sources in disturbing the proper functioning of the DNS protocol.

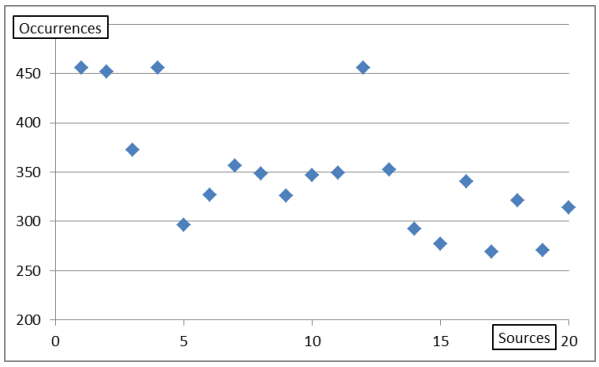


Figure 8. Occurrences of different sources

VII. CONCLUSIONS AND FUTURE WORK

The solution presented here, builds upon the existing solutions that collect statistical information regarding DNS services, by adding the ability to detect and control security incidents in real time. It also adds the advantage of operating in a distributed way, allowing the exchange of information between cooperating probes, and the reinforcement of its own security, even before it is threatened.

Currently, the solution presented does not allow the processing of addresses in the IPv6 format. The technical aspects that led to this situation are linked to the need to optimize the performance of the data recorder application making it possible to store the data from all consultations. Nevertheless, all queries made to IPv6 addresses are contained in this solution (AAAA types).

We are also working on extending the data correlation capabilities of the system by adding information collected from other sources (intrusion detection systems for instance). We anticipate that this could be a valuable approach to reduce considerably the number of false positives and negatives [16].

REFERENCES

[1] P. Vixie, “DNS Complexity”, ACM Queue vol. 5, no. 3, April 2007.

[2] D. Wessels, "A Recent DNS Survey", DNS-OARC, November 2007.

[3] Dave Piscitello, “Conficker Summary and Review”, ICANN, May 2010.

[4] SQLDNS website, [http://home.tiscali.cz:8080/~cz210552/sqldns.html]. Last accessed on 17 November 2010.

[5] BIND website, [http://www.isc.org/products/BIND]. Last accessed on 17 November 2010.

[6] Bojan Zdrnja, “Security Monitoring of DNS traffic”, May 2006.

[7] Paul Vixie, D. Wessels, ”DNSCAP – DNS traffic capture utility”, CAIDA Workshop, July 2007.

[8] Duane Wessels, "Whats New with DSC", DNS-OARC, November 2007.

[9] Lawrence Berkeley National Laboratory. Tcpdump website http://www.tcpdump.org.

[10] John Kristoff, "An Automated Incident Response System Using BIND Query Logs", June 2006.

[11] MySQL website – (Open Source Database), [http://www.mysql .com]. Last accessed on 17 November 2010.

[12] IP FILTER – TCP/IP Firewall/NAT Software, [http://coombs.anu.edu.au/~avalon]. Last accessed on 17 November 2010.

[13] P. Saint-Andre, Ed., Extensible Messaging and Presence Protocol (XMPP): Core, RFC 3920, 2004.

[14] João Afonso, Edmundo Monteiro, "Development of an Integrated Solution for Intrusion Detection: A Model Based on Data Correlation", in Proc. of the IEEE ICNS'06, International Conference on Networking and Services - ICNS'06, Silicon Valley, USA, July 2006.

[15] Tshark website – The Wireshark Network Analyzer, [http://www.wireshark.org]. Last accessed on 17 November 2010.

[16] João Afonso, Pedro Veiga, “Protecting the DNS Infrastructure of a Top Level Domain: Real-Time monitoring with Network Sensors”, WSNS 2008, 4th IEEE – International Workshop on Wireless and Sensor Networks Security, Atlanta, USA, 29 September – 2 October 2008.

 ISBN:978-1-61208-113-7

Design Experience with Routing SW and Related Applications

Miroslav Sveda
Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic
e-mail: sveda@fit.vutbr.cz

***Abstract*—This paper deals with the current software architectures for intermediate systems for Intranet and small-range wireless interconnection using case studies founded on real-world applications. The approach demonstrates another contribution to network convergence in interconnecting software architecture development, which stems from a design experience based on industrial network applications and on metropolitan networking. The first case study focuses on IEEE 1451 family of standards that provides a design framework for creating applications based not only on IP/Ethernet profile but also on ZigBee. Next case study explores how security and safety properties of Intranets can be verified under every network configuration using model checking. The contribution of the paper consists of a new method to network convergence and network modeling in software architecture development.**

Keywords-network architecture, sensor networks, intranets, validation of network configuration

I. INTRODUCTION

This paper focuses on software architectures for intermediate system control plane in frame of Intranet and ZigBee, and then presents new contributions to network convergence and network modeling in software architecture development. To facilitate comprehensible wording, the beginnings of this section and two following subsections restate basic, standard-based terminology used in the following text.

According to the ISO Open Systems Interconnection (ISO-OSI) vocabulary, two or more sub-networks are interconnected using equipment called as intermediate system whose primary function is to relay selectively information from one sub-network to another and to perform protocol conversion where necessary. A bridge or a router provides the means for interconnecting two physically distinct networks, which differ occasionally in two or three lower layers respectively. The bridge converts frames with consistent addressing schemes at the data-link layer, or medium access and control (MAC) sub-layer, while the router deals with packets at the network layer. Lower layers of these intermediate systems are implemented according to the proper architectures of interconnected networks. When sub-networks differ in their higher layer protocols, especially in the application layer, or when the communication functions of the bottom three layers are not sufficient for coupling, the intermediate system, called in this case as gateway, contains all layers of the networks involved and converts application messages between appropriate formats.

An intermediate system represents typically a node that belongs simultaneously to two or more interconnected networks. The backbone network interconnects more intermediate systems that enable to access different sub-networks. If two segments of a network are interconnected through another network, the technique called as tunneling enables to transfer protocol data units of the end segments nested in the proper protocol data units of the interconnecting network.

The next section corroborates the basic concepts of supporting resources, namely (1) IP routers as the most important means forming the Internet, (2) industrial network couplers that enable to create hierarchical communication systems as a basis of various -- not only industrial -- applications, and (3) design experience collected by our team in this domain, which influence unsurprisingly the current research.

Section III. dealing with network convergence aims at Ethernet and IP-based industrial networking that offer an application development environment compatible with common TCP/IP setting. It stems from IEEE 1451 family of standards and provides a design framework for creating applications based not only on TCP/IP/Ethernet profile but also on ZigBee. The second part of this section reviews the first case study based on an application dealing with pressure and temperature measurement and safety and security management along gas pipes.

In section IV., the presented network modeling approach provides a unifying model suitable for description of relevant aspects of real IP computer networks including dynamic routing and filtering. The rest of this section reviews the second case study based on an application exploring how security and safety

 ISBN:978-1-61208-113-7

properties can be verified under every network configuration using model checking.

II. STATE OF THE ART

A. IP routers

Internet/Intranet router architectures have experienced three generations [7]. The *first generaion* router architecture, sometimes called also as software router, which is based on a monolitic (or centralized) routing engine, appears just as a simple PC equiped with multiple line cards.

In a cluster-based architecture, often called as the *second generation*, the Routing Engine modules are distributed on several network communication cards that share an interconnection, usually through a system bus, to operation memory and processor on the control card.

Many current Internet routers, which can provide high speed switching capacity, are built with switching fabrics based on a Banyan or analogous self-routing topology[8]. Not only pure routing, but also additional network services have enriched router functionalities in the past few years, for Internet namely packet tagging, emulating application-level proxies, application-specific packet dropping, performance monitoring, intrusion detecting, and assorted filtering and firewalling. Nevertheless, the routing engine provides the essential part of router functionality. As a software component, the routing engine is used to control the router activities and to build the data forwarding table.

B. Industrial networks coupling

Contemporary industrial distributed computer-based systems encompass, at their lowest level, various wired or wireless digital actuator/sensor to controller connections. Those connections usually constitute the bottom segments of hierarchical communication systems that typically include higher-level fieldbus or Intranet backbones. Hence, the systems must comprise suitable interconnections of incident higher and lower fieldbus segments, which mediate top-down commands and bottom-up responses. While interconnecting devices for such wide-spread fieldbuses as CAN, Profibus, or WorldFIP are currently commercially available, some real-world applications can demand also to develop various couplers either dedicated to special-purpose protocols or fitting particular operational requirements, see [12].

The following taxonomy of industrial communication and/or control network (ICN) interconnections covers both the network topology of an interconnected system and the structure of its intermediate system, which is often called in the industrial domain as *coupler*. On the other hand, the term gateway sometimes denotes an accessory connecting PC or a terminal to an ICN. For this paper, the expression "gateway" preserves its original meaning according to ISO-OSI terminology as discussed above.

The first item to be classed appears the level ordering of interconnected networks. A peer-to-peer structure occurs when two or more interconnected networks interchange commands and responses through a bus coupler in both directions so that no one of the ICNs can be distinguished as a higher level. If two interconnected ICNs arise hierarchically ordered, the master/slaves configuration appears usual at least for the lower-level network.

The second classification viewpoint stems from the protocol profiles involved. In this case, the standard taxonomy using the general terminology mentioned above can be employed: bridge, router, and gateway. Also, the tunneling and backbone networks can be distinguished in a standard manner.

The next, refining items to be classed include internal logical architectures of the coupler, such as source or adaptive routing scheme, routing and relaying algorithms, and operating system services deployed.

C. Design backgrounds

We launched our coupling development initiatives in the Fieldbus and Internet domains almost concurrently, see [10] and [3]. Fieldbus coupling was studied by our research team originally from the viewpoint of network architecture of low-level fieldbuses [10][11]. Next interest was focused on real-world applications based on network coupling, such as data acquissition appliance [9], or wireless smart sensors [14]. And also, the role of Ethernet and TCP/IP attracted our attention as a means of network convergence [2][12].

The other branch of our network interconnection initiative covers IP routing. In this case we launched with software router design based on a simple Unix machine [3] and with creation of a routing domain for academic metropolitan networking [5]. The current research initiatives deal with the high-speed IP6 router for optical networks [16], and with modeling of dynamically routed IP networks and exploration of their properties such as reachability-based safety and security [6].

III. NETWORK CONVERGENCE

This section deals with network convergence aiming at Ethernet and IP-based industrial networking that offer an application development environment compatible with the common TCP/IP setting. It stems

 ISBN:978-1-61208-113-7

from IEEE 1451 family of standards, mentioned in the subsection 3.2 and provides a design framework for creating applications based not only on TCP/IP/Ethernet profile but also on ZigBee. The last part of this section reviews the first case study based on an appication dealing with pressure and temperature measurement and safety and security management along gas pipes.

A. IP over Ethernet profile

The attractiveness of Ethernet as an industrial communication bus is constantly increasing. However the original concept of the Ethernet, which was developed during seventies of the last century as communication technology for office applications, has to face some issues specific for industrial applications. The concept of the Ethernet proved to be very successful and encountered issues are being addressed by modifications and extensions of the most popular 10/100 BaseT standard. In fact, the switched Ethernet with constraint collision domains proved to be efficient real-time networking environment also for time-critical applications.

Similarly, IP networking support appears as a rapidly dominating tendency in current industrial system designs. Namely, when layered over a real-time concerning data-link protocol, it seams as a best choice for future applications because of a simple interfacing within the Internet.

B. IEEE 1451 profile

The design framework, presented in this paper as a flexible design environment kernel, is rooted in the IEEE 1451.1 standard specifying smart transducer interface architecture. That standard provides an object-oriented information model targeting software-based, network independent, transducer application environments. The framework enables to unify interconnections of embedded system components through wireless networks and Ethernet-based intranets, which are replacing various special-purpose Fieldbuses in industrial applications [12].

The IEEE 1451 package consists of the family of standards for a networked smart transducer interface. The 1451.1 software architecture provides three models of the transducer device environment: (i) the object model of a network capable application processor (NCAP), which is the object-oriented embodiment of a smart networked device; (ii) the data model, which specifies information encoding rules for transmitting information across both local and remote object interfaces; and (iii) the network communication model, which supports client/server and publish/subscribe paradigms for communicating information between NCAPs. The standard defines a network and transducer hardware neutral environment in which a concrete sensor/actuator application can be developed.

The object model definition encompasses the set of object classes, attributes, methods, and behaviors that specify a transducer and a network environment to which it may connect. This model uses block and base classes offering patterns for one Physical Block, one or more Transducer Blocks, Function Blocks, and Network Blocks. Each block class may include specific base classes from the model. The base classes include Parameters, Actions, Events, and Files, and provide component classes.

The Transducer Block abstracts all the capabilities of each transducer that is physically connected to the NCAP I/O system. During the device configuration phase, the description of what kind of sensors and actuators are connected to the system is read from the hardware device. The Transducer Block includes an I/O device driver style interface for communication with the hardware. The I/O interface includes methods for reading and writing to the transducer from the application-based Function Block using a standardized interface.

The Function Block provides a skeletal area in which to place application-specific code. The interface does not specify any restrictions on how an application is developed.

The Network Block abstracts all access to a network employing network-neutral, object-based programming interface supporting both client-server and publisher-subscriber patterns for configuration and data distribution.

C. ZigBee profile

The ZigBee/IEEE 802.15.4 protocol profile [1][15] is intended as a specification for low-powered wireless networks. ZigBee is a published specification set of higher level communication protocols designed to use small low power digital radios based on the IEEE 802.15.4 standard for wireless personal area networks. The document 802.15.4 specifies two lower layers: physical layer and medium access control sub-layer. The ZigBee Alliance builds on this foundation by providing a network layer and a framework for application layer, which includes application support sub-layer covering ZigBee device objects and manufacturer-defined application objects.

Responsibilities of the ZigBee network layer include mechanisms used to join and leave a network, to apply security to frames and to route frames to their intended destinations. In addition to discovery and maintenance of routes between devices including

 ISBN:978-1-61208-113-7

discovery of one-hop neighbors, it stores pertinent neighbor information. The ZigBee network layer supports star, tree and mesh topologies

The ZigBee application layer includes application support sub-layer, ZigBee device objects and manufacturer-defined application objects. The application support sub-layer maintains tables for binding, which is the ability to match two devices together based on their services and their needs, and forwards messages between bound devices.

D. Sensor network case study

This section describes a case study that demonstrates deployment of the introduced design concepts. The application deals with pressure and temperature measurement and safety and security management along gas pipes. The related implementation stems from the IEEE 1451.1 model with Internet and the IEEE 1451.5 wireless communication based on ZigBee running over the IEEE 802.15.4.

The interconnection of TCP/IP and ZigBee is depicted on Figure 1. It provides an interface between ZigBee and IP devices through an abstracted interface on IP side. Each wireless sensor group is supported by its controller providing Internet-based clients with secure and efficient access to application-related services over the associated part of gas pipes. In this case, clients communi-cate to controllers using a messaging protocol based on client-server and subscribe-publish patterns employing 1451.1 Network Block functions. A typical configuration includes a set of sensors generating pressure and temperature values for the related controller that computes profiles and checks limits for users of those or derived values. When a limit is reached, the safety procedure closes valves in charge depending on safety service specifications.

Security configurations can follow in this case the tiered network architecture: (1) To keep the system maintenance simple, all wireless communication uses standard ZigBee hop-by-hop encryption based on single network-wide key because separate pressure and/or temperature values, which can be even-dropped, appear useless without the overall context; (2) Security in frame of Intranet subnets stems from current virtual private network concepts such that the communicating couples utilize ciphered channels based on tunneling between each client and a group of safety valve controllers -- the tunnels are created with the support of associated authentications of each client.

The example network configuration, see Figure 2., comprises several groups of wireless pressure and temperature sensors with safety valve controllers as base stations connected to wired intranets that dedicated clients can access effectively through Internet. The WWW server supports each sensor group by an active web page with Java applets that, after downloading, provide clients with transparent and efficient access to pressure and temperature measurement services through controllers. Controllers offer clients not only secure access to measurement services over systems of gas pipes, but also communicate to each other and cooperate so that the system can resolve safety and security-critical situati ons by shutting off some of the valves.

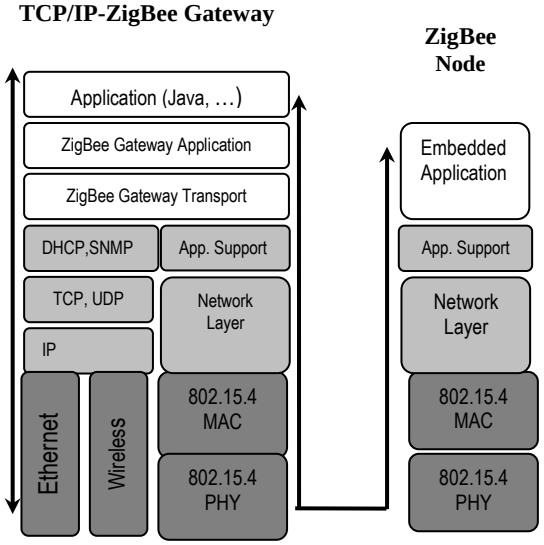


Figure 1: Network gateway.

Each controller communicates wirelessly with its sensors through 1451.5 interfaces by proper communication protocol. In the discussed case the proposed P1451.5-ZigBee, which means ZigBee over IEEE 802.15.4, protocol was selected because it fits application requirements, namely those dealing with power consumption, response timing, and management. The subscriber-publisher style of communication, which in this applica-tion covers primarily distribution of measured data, but also distribution of group configuration commands, employs IP multicasting. All regular clients wishing to receive messages from a controller, which is joined with an IP multicast address of class D, register themselves to this group using IGMP. After that, when this controller generates a message by Block function publish, this message is delivered to all members of this class D group, without unnecessary replications.

The WWW server supports each sensor group by an active web page with Java applets that, after downloading, provide clients with transparent and efficient access to pressure and temperature

measurement services through controllers. Controllers provide clients not only with secure access to measurement services over systems of gas pipes, but also communicate to each other and cooperate so that the system can resolve safety and security-critical situations by shutting off some of the valves.

Each wireless sensor group is supported by its controller providing Internet-based clients with secure and efficient access to application-related services over the associated part of gas pipes. In this case, clients communicate to controllers using a messaging protocol based on client-server and subscriber-publisher patterns employing 1451.1 Network Block functions. A typical configuration includes a set of sensors generating pressure and temperature values for the related controller that computes profiles and checks limits for users of those or derived values. When a limit is reached, the safety procedure, which is derived from the fail-stop model, closes valves in charge depending on safety service specifications.

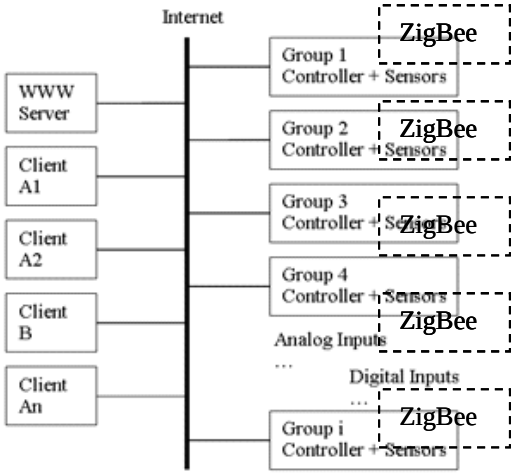


Figure 2: Example network configuration.

IV. DYNAMIC NETWORK BEHAVIOR

The current goals of our research in frame of Internet-level routed networks consist of i) creation of a unifying model suitable for description of relevant aspects of real computer networks including routing information, ACLs (access control lists), NAT (network address translation), dynamic routing policy; and ii) delivering methods for automated verification of dependable properties (e.g., availability, security, survivability). The unique added value of the project is to specifically merge the research on formal methods with the research on network security to devise a new method for network security verification.

A. *Dynamic network model*

The recent work has focused on studying models and analysis techniques based on simulation and network monitoring [6]. These models, nevertheless, do not take into account routing and packet filtering despite the fact that these aspects may significantly influence the traffic coverage observed in the network. The intensive research needs to be done in order to find new models that would include dynamic view on the network.

Similarly to hardware and software analysis based on simulation, the network simulation methods are useful mainly to observe properties given by regular behavior of the system. Simulation techniques are incompetent in catching "what if" cases that occur rarely in the system. However, the real world systems inevitably exhibit also the unusual behavior. The use of formal methods is better suited for checking those situations to uncover hidden problems.

The dynamics of current network models is most often limited to changes of actual data in time. The other dimension of dynamics of routed networks comes from dynamic routing protocols and topology changes based on the availability of links and link parameters, e.g., reliability, bandwidth or load. The anticipated project results characterize a novel approach in the area of network traffic analysis.

B. *Dynamic network case study*

The recent work has focused on studying models and analysis techniques based on simulation and network monitoring [6][13]. These models, nevertheless, do not take into account routing and packet filtering despite the fact that these aspects may significantly influence the traffic coverage observed in the network. The intensive research needs to be done in order to find new models that would include dynamic view on the network.

In our work we explore how security and safety properties can be verified under every network configuration using model checking [4]. The model checking is a technique that explores all reachable states and verifies if the specified properties are satisfied over each possible path to those states. Model checking requires specification of a model and properties to be verified. In our case, the model of network consists of hosts, links, routing information and ACLs. The network security-type properties can be expressed in the form of modal logics formulae as constraints over states and execution paths. If those formulas are not satisfied, the model checker generates

 ISBN:978-1-61208-113-7

a counterexample that reveals a state of the network that violates the specification. If the formulas are satisfied, it means, that the property is valid in every state of the systems, see more detail in [13].

V. CONCLUSIONS

The paper discusses software architectures for intermediate system's control planes belonging to Intranets and Fieldbuses by two case studies derived from genuine implementations. The interest is focused both on network convergence and on network modeling in application architectures development. The applied solutions stem from design experience both with industrial network appliances and metropolitan networking. The first case study focuses on IEEE 1451 family of standards that provides a useful design framework for creating applications based not only on IP/Ethernet profile but also on ZigBee over IEEE 802.15.4. Next case study explores how security and safety properties of interconnected intranets can be verified under every network configuration using model checking.

Note that several various methods may fit modeling and analysis of the properties in the domains of interest. Most often, the combination of several methods leads to better results. The emphasis of the project's research is put on the formal verification methods, but other methods are certainly worthwhile to explore as well. The other methods may be orthogonal with formal verification, or they may support the formal methods.

In particular, monitoring may provide a fruitful data for classification and definition of security-related properties based on the real traffic. Modeling and simulation serve as a useful tool to specify and replay possible dangerous scenarios found by the formal verification. Therefore, simulators and monitors can efficiently support network-wide analysis namely during the design and development.

ACKNOWLEDGEMENTS

This project has been partially supported by the BUT FIT grant FIT-10-S-2, the research plan MSM0021630528 and through the grant no. GACR 102/08/1429. Also, the author was partially supported by the grant no. FR-TI1/037 of Ministry of Industry and Trade.

The author acknowledges contributions to this work by his colleagues (alphabetically) Petr Matousek, Ondrej Rysavy, Jaroslav Rab, Roman Trchalik, Radimir Vrba and Frantisek Zezulka.

REFERENCES

[1] P. Baronti, P. Pillai, V. Chook, S. Chessa, A. Gotta, and Y. Fun Hu, "Wireless sensor networks: A survey on the state of the art and the 802.15.4 and ZigBee standards," Computer communications, Vol.30, 2007, pp.1655-1695.

[2] P. Cach, P. Fiedler, M. Sveda, M. Prokop, and M. Wagner, "A Sensor with Embedded Ethernet," In WSEAS Transactions on Circuits, Iss.1, Vol.2, 2003, pp.213-215.

[3] I. Cernohlavek, J. Novotny, V. Slama, V. Zahorik, and M. Sveda, "Open-Box Routers with Academic Metropolitan Networking," Technical Report, Brno University of Technology and Masaryk University, Brno, 1994.

[4] E.M. Clarke, O. Grumberg, and D.A. Peled, Model Checking, MIT Press, Boston, 1999.

[5] L. Kania, S. Smolik, M. Sveda, and V. Zahorik, "The Brno Academic Computer Network and its Future Development," In Proceedings INVEX-CCT'95, BVV Press, Brno, 1995, pp.1-5.

[6] P. Matousek, J. Rab, O. Rysavy, and M. Sveda, "A Formal Model for Network-wide Security Analysis," In Proceeding of the 15th IEEE International Symposium and Workshop on the Engineering of Computer-based Systems, Belfast, GB, IEEE Computer Society, Los Alamitos, 2008, pp.171 181.

[7] K. Nguyen and B. Jaumard, "Routing Engine Architecture for Next Generation Routers: Evolutional Trends," In International Journal of Network Protocols and Algorithms, Vol.1, No.1, Macrothink Institute, Las Vegas, Nevada, 2009, pp.62-85.

[8] A. Nucci and K. Papagiannaki, Design, Measurement and Management of Large-Scale IP Networks: Bridging the Gap between Theory and Practice, Cambridge University Press, New York, 2009.

[9] O. Sajdl, Z. Bradac, R. Vrba and M. Sveda, "Data Acquisition System Exploiting Bluetooth Technology, In WSEAS Transactions on Circuits," Iss.1, Vol.2, 2003, pp.117-119.

[10] M. Sveda, "Routers and Bridges for Small Area Network Interconnection," In Computers in Industry, Vol.22, No.1, Elsevier Science, Amsterdam, NL, 1993, pp.25 29.

[11] M. Sveda, R. Vrba, and F. Zezulka, "Coupling Architectures for Low-Level Fieldbuses," In Proceedings 7th IEEE ECBS Conference, Edinburgh, Scotland, IEEE Comp. Soc., 2000m pp.148-155.

[12] M. Sveda, P. Benes, R. Vrba, and F. Zezulka, "Introduction to Industrial Sensor Networking," Book Chapter in M. Ilyas, I. Mahgoub (Eds.): Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems, CRC Press LLC, Boca Raton, FL, 2005, pp.10.1-10.24.

[13] M. Sveda, O. Rysavy, P. Matousek, and J. Rab, "An Approach for Automated Network-Wide Security Analysis," In: Proceedings of the Ninth International Conference on Networks ICN 2010, Les Menuires, FR, IARIA, IEEE CS, 2010, pp. 294-299.

[14] R. Vrba, O. Sajdl, R. Kuchta, and M. Sveda, "Wireless Smart Sensor Network System," In Proceedings of the Joint International Systems Engineering Conference (ICSE) and The International Council on Systems Engineering (INCOSE), Las Vegas, Nevada, 2004.

[15] ZigBee, 2006. ZigBee Specification. ZigBee Alliance Board of Directors Website http://www.zigbee.org/.

[16] M. Kosek and J. Korenek, "FlowContext: Flexible Platform for Multigigabit Stateful Packet Processing," In: 2007 International Conference on Field Programmable Logic and Applications, Los Alamitos, US, IEEE CS, 2007, pp. 804-807.

 ISBN:978-1-61208-113-7

Adaptive Load Balanced Routing for 2-Dilated Flattened Butterfly Switching Network

Ajithkumar Thamarakuzhi John A. Chandy

Department of Electrical & Computer Engineering
University of Connecticut, Storrs, CT USA 06269–2157
{ajt06010,chandy}@engr.uconn.edu

***Abstract*—High-radix networks such as folded-Clos outperform other low radix networks in terms of cost and latency. The 2-dilated flattened butterfly (2DFB) network is a nonblocking high-radix network with better path diversity and reduced diameter compared to the folded-Clos network. In this paper, we introduce an adaptive load balanced routing algorithm that is designed to exploit all the positive topological properties of a 2DFB network. The proposed algorithm achieves load balance by allowing one non minimal forwarding in each dimension in case of network congestion. This algorithm provides high throughput on adversarial traffic patterns and provides better latency on benign traffic patterns. We have compared the performance of our algorithm on a 2DFB network with an Adaptive Clos algorithm on a folded-Clos network and a Minimal routing algorithm on a 2DFB network for different traffic patterns. We observed that 2DFB network with the proposed algorithm provides the same throughput with reduced latency compared to the folded-Clos network with an Adaptive Clos algorithm for all the traffic patterns.**

***Keywords*-Routing; adaptive; switching architecture;**

I. INTRODUCTION

High performance computing on distributed memory parallel processing systems such as clusters are very dependent on communication between processing nodes. As a result, the interconnection network that connects these nodes is a critical part of the performance of the system. For the past few decades, we have seen improving performance of processors and memory systems. In order to keep up with this, the network switch performance must also improve. The study of interconnection networks has a long history and a large number of network topologies and routing algorithms have been studied by researchers. Among these networks, hypercube [1] and Clos [2] (or its derivatives) are the most popular networks.

The technological progress in modern ASICs has led to the availability of routers with high bandwidth in the range of Tb/s. The improved pin bandwidth of these routers can be efficiently used to construct high-radix network topologies. Recent work has shown that high-radix network outperforms corresponding low-radix network in terms of cost and latency. Folded-Clos and flattened butterfly [3] are two topologies which can take advantage of the high-radix routers. The 2-dilated flattened butterfly (2DFB) is a nonblocking version of a flattened butterfly network. In previous work [4], [5], we have introduced the 2DFB network and proved its nonblocking behavior and shown the implementation of a 2DFB network switch using the NetFPGA platform. In this paper we propose an adaptive load balanced algorithm for 2DFB and observe its performance for different traffic patterns.

A routing algorithm can be considered as optimal if it provides low latency on local traffic and high throughput on adversarial traffic. Most algorithms must compromise one goal in order to achieve the other. Minimal routing, which always chooses the shortest path for each packet, provides minimum latency for local and benign traffic. However, it provides non acceptable latency for adversarial traffic due to load imbalance. In order to improve the throughput in adversarial traffic, the routing algorithm should balance the load by sending some fraction of packets over non-minimal paths.

Researchers have been trying to address the issue of providing high worst-case performance while preserving locality. Valiant's randomized algorithm [6] gives good performance in worst case traffic but very poor performance for local traffic in terms of latency. Minimal adaptive routing [7] [8] suffers from global load imbalance. GOAL is a load balanced adaptive routing algorithm designed for a torus network [9]. It provides better load balance with improved performance for local traffic. It achieved 58% throughput of the Minimal algorithm on nearest neighbor traffic for a torus network. Adaptive Clos [10] is an adaptive routing algorithm designed for Clos network which provides optimum performance for a high-radix Clos network. The adaptive routing algorithm that we propose in this paper is designed for a 2DFB network and it balances the load efficiently by allowing one non-minimal forwarding in each dimension in case of traffic congestion. It senses the traffic congestion from the packet queue. We observed the performance of this algorithm for local traffic and it has reduced latency than a Clos network with the Adaptive Clos algorithm.

The remainder of the paper is organized as follows.

 ISBN:978-1-61208-113-7

In Section II we briefly describe 2DFB and few of its topological properties. Section III describes the proposed adaptive load balanced algorithm for 2DFB network. In Section IV we present the simulation results and we conclude in Section V.

II. BACKGROUND

In this section we describe the 2DFB network [4] and its topological properties.

A. 2-dilated flattened butterfly structure

A 2DFB network is derived from a flattened butterfly structure [3] by either duplicating all the interconnecting links between the switching elements or replacing it with links of double bandwidth. Links between the end-terminals and switching elements remain the same. A 2DFB is composed of N/k routers of radix k'=$n(k-1)+1$ where N is the number of end-terminals in the network, n is the number of columns in a butterfly network, k is the number of end-terminals connected to each router and the radix(k') is the number of external ports associated with each router. The routers are connected by channels in $n'=n-1$ dimensions. In each dimension d, from 1 to n', router i is connected to each router j given by

$$j = i + [m - (\lfloor \frac{i}{k^{d-1}} \rfloor \mod k)]k^{d-1} \qquad (1)$$

for m from 0 to $k-1$, where the connection from i to itself is omitted. For example a 4-ary 2-dimensional 2DFB for N=64 is shown in Fig. 1.

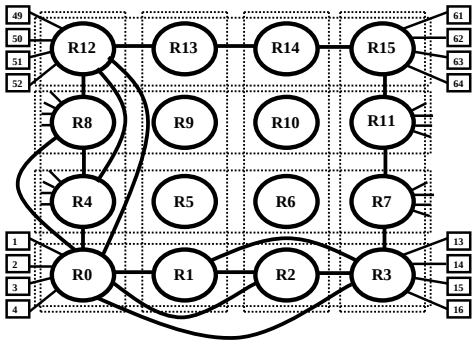


Figure 1. 4-ary flattened butterfly structure

As we can see in the Fig. 1, each switching element is connected to k end-terminals (here k=4). k switching elements in each row are interconnected and it can be considered as a 1-dimensional system. A 1-dimensional system is a fully connected ring structure with each link having double bandwidth. Its bisectional bandwidth is $N/2$ where N is the total number of end-terminal connected to the 1-dimensional system. In [4] we have proved that in a 1-dimensional 2DFB system any routing permutation can be performed without conflict using a maximum of two links. Higher dimensional 2DFB systems are constructed by combining 1-dimensional systems as shown in Fig. 1. For a k-ary d-dimensional [d=$(log_k N)-1$] 2DFB system with a network size (N) of power of k, the bisection bandwidth is $((k^2/2)(k^{d-1}))$ which is equal to $N/2$ (same as that of a hypercube network). Therefore, a properly designed routing algorithm can route any permutation without conflict by making use of a maximum of $2d$ hops (2 hops in each dimension).

B. Network diameter

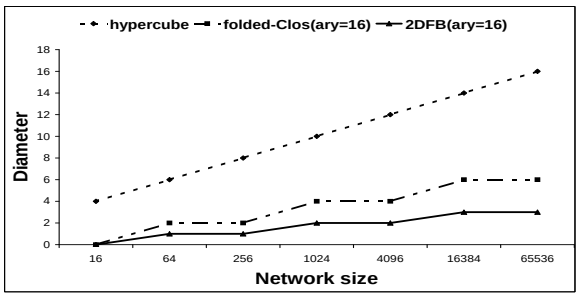


Figure 2. Network diameter

Network diameter is a measure of shortest distance between the source and destination nodes. Since high priority traffic can be routed through this shortest path, the network diameter plays an important role in a multi-processor communication system. A comparison of network diameter of 2DFB with other topologies for different network size is shown in Fig. 2. The diameter of a hypercube network is $log_2 N$, the diameter of a k-ary folded-Clos network is $2\{\lceil(log_k N)\rceil - 1\}$ and the diameter of a k-ary 2DFB is $\lceil(log_k N)\rceil - 1$. As we can observe, 2DFB has the smallest network diameter compared to other network topologies.

C. Number of hops

Message latency in a network is proportional to the number of hops required for routing the message. Fig. 3 represents the number of hops needed for routing the message for different network topologies with varying network sizes. Number of hops required in 2DFB is not same as the network diameter for all source-destination pair. For example in Fig. 1 if end-terminals 1,2,3 and 4 are sending messages to end-terminals 5,6,7 and 8 respectively with full bandwidth, then only messages from terminal 1 and 2 can be routed through the direct link between R_0 and R_1 and the messages from 3 and 4 should be routed through R_2 or R_3. In this case the number of hops required in the worst case is 2. In higher dimension 2DFB, in worst case, 2 hops

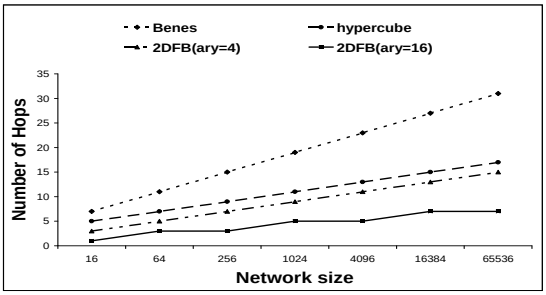


Figure 3. Number of hops needed for routing

are required for routing the message in each dimension. The dimension of a k-ary 2DFB is $\lceil(log_k N)\rceil - 1$. So the number of hops required (worst case) to complete any routing request in a k-ary 2DFB for $k \gg 2$ is $2\{\lceil(log_k N)\rceil - 1\}$. For $k = 2$, 2DFB becomes a normal hypercube structure with 2 end-terminals connected to each switching element and with interconnecting links of double bandwidth. In this case the number of hops required is $log_2(N/2)$. In [11] it is shown that the number of hops required in a hypercube network is $log_2 N$. The number of hops required for a k-ary folded-Clos network is $2\{\lceil(log_k N)\rceil - 1\}$. From the comparison we can see that the number of hops required for a k-ary 2DFB in the worst case is the same as that of a k-ary folded-Clos network. Unlike folded-Clos, in 2DFB the number of hops required is not same for all the source-destination pair. Large amount of source-destination pair need only one hop to traverse in one dimension. So the average number of hops in a 2DFB will be always less than that of the corresponding folded-Clos network. Thus it is clear that k-ary 2DFB provides better message latency than corresponding folded-Clos network.

D. Cost Analysis

A key determinant of the effectiveness of a network topology is the cost of the network relative to the performance it delivers. Cost of the network is decided by cost of routers and links. The number of switching elements and links required to implement a 2DFB network is less than other nonblocking networks such as folded-Clos and hypercube and therefore the implementation cost of a 2DFB network will be lesser than other nonblocking networks [4].

III. Routing Algorithm

The proposed routing algorithm is designed to explore the topological properties of a 2DFB network. A 2DFB network is similar to a k-ary generalized hypercube (GHC) except that in a 2DFB k end-terminals are connected to each switching element. A 2DFB can be considered as a k-way bristled 2-dilated GHC. If r is the dimension of a k-ary flattened butterfly, then there will be k^r nodes(switching elements) in the system and each node can be represented using a r-digit number, i.e. any node $x = x_{r-1}...x_i...x_0$ where $x_i \in [0, k-1]$. In a 2DFB network any two nodes, whose numbers differ only in the ith digit, are joined by a duplex channel and it is known as the ith dimension channel. Thus by comparing the r bit number associated to the current switching element and the destination switching element, one can find out the set of dimensions in which forwarding of the packet is required. Every node contains $(k-1)$ channels in each dimension.

The proposed routing algorithm has two phases of operation, minimal forwarding phase and non-minimal forwarding phase. In the minimal phase, the algorithm considers the set of dimensions in which forwarding is required and it adaptively selects the dimension if the direct link in the selected dimension is ready to use. We are using a sequential allocation method in our algorithm which gives maximum performance. If no direct link is available in any of the selected dimension in the minimal phase, then the algorithm will turn in to non-minimal phase of operation.

In non-minimal phase the algorithm will consider all selected dimension and adaptively check the availability of any of the non-minimal link in the selected dimension. If it finds any available non-minimal link, the packet will be forwarded to that link. We constrain this non-minimal forwarding by adding one bit flag in the header of each packet and we call this flag as the priority flag. The algorithm allows only one non-minimal forwarding in each dimension. If the switching element sees that the priority flag is set for the received packet, then that packet will be sent to a minimal direct link even though all minimal output queue have packets more than the threshold level. In the next cycle some portion of the traffic coming from the other switches will be adaptively rerouted to any non-minimal link which will reduce the traffic congestion. Thus, by the combined use of minimal and non-minimal phase of operation the algorithm will balance the load efficiently and it will reach the steady state within a few iterations.

The algorithm always gives priority to the minimal forwarding and therefore for local traffic and benign traffic, the performance of this algorithm will be very close to the minimal routing. With the worst case traffic the algorithm will use at most two links per dimension. In the worst case also a fraction of traffic is routed through direct links. So the average latency will be still less than that of a Adaptive Clos algorithm in a Clos network.

A. Algorithms used for comparison

We have selected Minimal and Adaptive Clos routing algorithms for the performance comparison with our proposed adaptive algorithm. The Minimal algorithm will always route packets in the shortest path. Adaptive Clos routing have

 ISBN:978-1-61208-113-7

forward and backward phases. In the forward phase any of the output queue in the forward path is adaptively selected by considering the number of packets in each output queue. In the reverse phase routing is deterministic as there exists only a single path to the destination. The Minimal routing algorithm is implemented in a 2DFB and the Adaptive Clos routing is implemented in Clos networks.

B. Terminologies used in the algorithm

The proposed adaptive routing algorithm is shown in the Algorithm 1. A one bit flag is added to the header of each packet to indicate the switching priority and it is represented as h_1. An output port is selected by considering the number of packets already in queue in the corresponding output queue. The port is selected if the number of packets in the output queue is less than the threshold value T_h. The preferred output ports are also decided by comparing the r digit representation of the current switching element and the destination switching element, where r is the dimension of the network. r digit representation of current switching element and destination switching element is represented as $s_d[r]$ and $d_d[r]$ respectively. $dimsel$ is a pointer to the selected dimension and P_s is a flag indicating whether a port is selected or not.

IV. Results

We have modeled 2DFB and folded-Clos networks for different network sizes using the OMNeT++ simulation library [12]. These topologies are implemented using interconnecting links of 2 Gb/s bandwidth. All the end-terminals are sending packets with a maximum bandwidth of 1 Gb/s. We have used a packet size of 121 bytes. Higher size packets are also following the same trend. The default OMNeT switch model was modified in order to include a 2 Gb/s channel. We have compared the throughput and latency of these network topologies for different traffic patterns. We assume that the data transmission through the network is permutation type - i.e. a unique source and destination are assigned to any data element and the elements are permuted upon transmission. We have selected three traffic patterns to consider the best case and worst case scenario of 2DFB topology which are named as below.

1) $Benign$: In a 2DFB structure each switching element is connected to $k-1$ switching elements using direct links in each dimension. In $benign$ traffic pattern all the traffic can be routed through these directed links, that is in this pattern the number of hops required for the routing of any packet will be equal to the diameter of the 2DFB network. In this pattern each pair of end-terminals connected to a switching element will be sending traffic to different directly connected switching elements. 2DFB provides minimum latency for $benign$ traffic pattern.

2) $Adversarial$: In this traffic pattern all the end-terminals connected to a switching element S_i will be sending traffic to end-terminals which are connected to another single switching element S_{i+j}. If this pattern is used in a 2DFB only two end-terminals which are connected to a switching element can send traffic through the direct link. All the other $k-2$ end-terminals should send traffic through indirect links. 2DFB provides worst case latency for $adversarial$ traffic pattern.

3) $Random$: In this pattern destination terminals are selected randomly. Latency provided by 2DFB for this pattern will be between that of $benign$ and $adversarial$ patterns.

A. Throughput comparison

We have compared the average throughput of a 8-ary 1-dimensional network with a network size of 64 for three different routing algorithms, Minimal, Adaptive Clos and our proposed algorithm which is named as Adaptive 2DFB. Minimal and Adaptive 2DFB algorithms are implemented over a 2DFB network. Adaptive Clos routing algorithm is implemented over a Clos network with the same size.

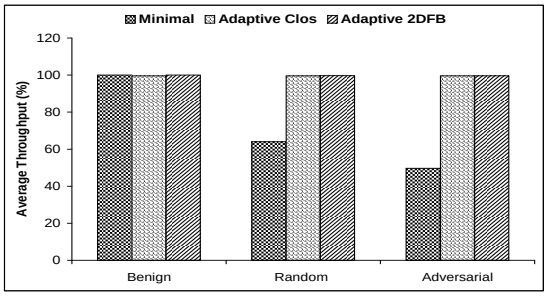


Figure 4. Throughput comparison of 1-dimensional networks

The throughput comparison is done for three different traffic patters as mentioned before. As shown in Fig. 4, like the Clos network, 2DFB also provides throughput which is very close to 100% for all the given traffic patterns. The Minimal algorithm provides 100% throughput only for benign traffic and it provides 50% throughput for adversarial traffic pattern. This shows the benefit of our adaptive algorithm as it is able to maintain high throughput in both adversarial and benign traffic patterns.

We have also compared the average throughput of a 8-ary, 2-dimensional 2DFB with 8-ary, 2-dimensional Clos network. Both of the network have a network size of 512. The throughput comparison is shown in Fig. 5. Two dimensional network also provides similar average throughput as one dimensional network.

B. End-to-end packet delay comparison

We have compared the average end-to-end packet delay of a 8-ary 1-dimensional and 2-dimensional networks for dif-

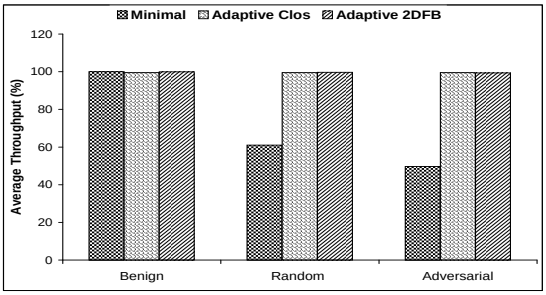


Figure 5. Throughput comparison of 2-dimensional networks

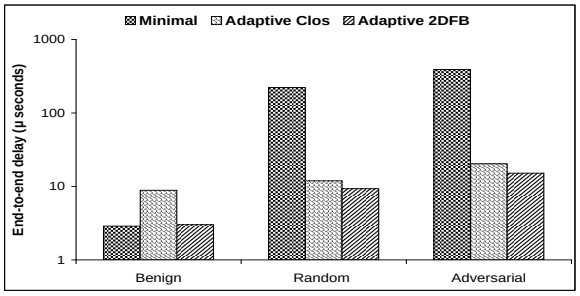


Figure 7. End-to-end packet delay comparison of 2-dimensional networks

ferent routing algorithms. Average end-to-end packet delay comparison of 8-ary 1-dimensional networks with a network size of 64 is shown in Fig. 6.

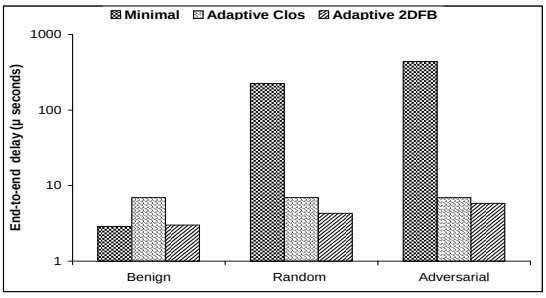


Figure 6. End-to-end packet delay comparison of 1-dimensional networks

In Fig. 6 we can notice that the average end-to-end packet delay of 2DFB for benign traffic pattern is less than that of adversarial traffic pattern. We can also notice that the average end-to-end packet delay of the Adaptive 2DFB algorithm is less than that of Adaptive Clos algorithm for all the traffic patterns. As would be expected, the Minimal algorithm shows poor load balancing and provides very high packet delay for adversarial traffic pattern compared to other algorithms.

Average end-to-end packet delay comparison of 8-ary, 2-dimensional networks with a network size of 512 is shown in Fig. 7. We can observe that 2-dimensional networks also follow the same trend as 1-dimensional networks. Both the 1-dimensional and 2-dimensional 2DFB networks with Adaptive 2DFB algorithm provide maximum end-to-end packet delay for adversarial traffic pattern. This maximum value is still less than corresponding Clos network with Adaptive Clos algorithm. Practical traffic patterns will be random in nature and the end-to-end packet delay of Adaptive 2DFB algorithm in a 2DFB network, for the random traffic pattern will be in between the end-to-end packet delay of benign and adversarial traffic patterns. This end-to-end packet delay comparison reveals the effectiveness of the proposed Adaptive 2DFB routing algorithm on 2DFB networks. This comparison also shows the benefit of the 2DFB architecture with respect to Clos in that maintains high throughput with lower latency costs than the more expensive Clos architecture.

V. Conclusion

In this paper, we have introduced an adaptive load balanced routing algorithm for 2DFB switching network. The proposed algorithm is designed to exploit the nonblocking property of 2DFB network. The algorithm also takes full advantage of the reduced diameter of 2DFB network. It provides better load balancing by allowing one non-minimal forwarding in each single dimension of 2DFB which is a 2-dilated fully connected ring structure. This algorithm also provides good performance for local and benign traffic by providing priority to the selection of direct links. We have compared the performance of the proposed algorithm running over a 2DFB with the Adaptive Clos algorithm running over a Clos network and Minimal routing algorithm over a 2DFB network, and we have observed that our algorithm provides reduced latency for all the traffic patterns while maintaining the same throughput of the Adaptive Clos algorithm. Thus, we conclude that the 2DFB with the proposed algorithm will be an optimal candidate for a high performance interconnection system with reduced cost.

Acknowledgment

This work was supported in part by a National Science Foundation High End Computing University Research Activity grant (award number CCF-0621448). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

Algorithm 1: Adaptive routing algorithm.

```
1  begin
2      if s_d == d_d then
3          P_s = 1
4          port = read (port for the end-terminal)
5          send(frame, port)
6      else
7          % Minimal forwarding phase
8          Set i= msb and db=0
9          dimsel= new int[r]
10         repeat
11             if s_d[i] == d_d[i] then
12                 goto deci
13             else
14                 *(dimsel+db)=i
15                 port = read (direct port)
16                 if h_1 == 1 then
17                     chk:if port == input port then
18                         db = db+1 and goto deci
19                     else
20                         set h_1 to 0 and P_s to 1
21                         send(frame, port)
22                         goto sel0
23                     end
24                 else
25                     if packets in queue <= T_h then
26                         goto chk
27                     else
28                         db=db+1 and goto deci
29                     end
30                 end
31             end
32             deci: i = i - 1
33         until i >= 0
34         sel0:if P_s == 0 then
35             % Non-minimal forwarding phase
36             s= ary-2
37             for b ← 0 to db do
38                 dims = *(dimsel+b)
39                 repeat
40                     port = (dims*(ary-1))+s
41                     if port == input port then
42                         goto decrement
43                     else
44                         if packets in queue <= T_h then
45                             Set h_1 to 1 and P_s to 1
46                             send(frame, port)
47                             break
48                         end
49                         goto decrement
50                     end
51                     decrement: s = s - 1
52                 until s >= 0
53             end
54         end
55     end
56 end
```

REFERENCES

[1] L. N. Bhuyan and D. P. Agrawal, "Generalized hypercube and hyperbus structures for a computer network," *IEEE Trans. Computers*, vol. 33, no. 4, pp. 323–333, 1984.

[2] C. Clos, "A study of non-blocking switching networks," *The Bell System Technical Journal*, vol. 32, pp. 406–424, Mar. 1953.

[3] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: A cost-efficient topology for high-radix networks," in *Proc. of the International Symposium on Computer Architecture (ISCA)*, pp. 126–137, June 2007.

[4] A. Thamarakuzhi and J. A. Chandy, "2-dilated flattened butterfly: A nonblocking switching network," in *International Conference on High Performance Switching and Routing (HPSR 2010)*, June 2010.

[5] A. Thamarakuzhi and J. A. Chandy, "Design and implementation of a nonblocking 2-dilated flattened butterfly switching network," in *IEEE Latin-American Conference on Communications 2010*, 2010.

[6] L. G. Valiant and G. J. Brebner, "Universal schemes for parallel communication," in *In Proc. of the ACM Symposium on the Theory of Computing*, pp. 263–277, 1981.

[7] L. Gravano, G. Pifarre, G. Pifarre, P. Berman, and J. Sanz, "Adaptive deadlock- and livelock-free routing with all minimal paths in torus networks," *IEEE Trans. on Parallel and Distributed Systems*, vol. 5, no. 12, pp. 1233–1252, 1994.

[8] D. Linder and J. Harden, "An adaptive and fault tolerant wormhole routing strategy for k-ary n-cubes," *ACM Trans. on Computer Systems*, vol. 40, no. 1, pp. 2–12, 1991.

[9] A. Singh, W. J. Dally, A. K. Gupta, and B. Towles, "Goal: A loadbalanced adaptive routing algorithm for torus networks," in *In Proc. of the International Symposium on Computer Architecture*, pp. 194–205, June 2003.

[10] J. Kim, W. J. Dally, and D. Abts, "Adaptive routing in high-radix clos network," in *In International Conference for High Performance Computing, Networking, Storage, and Analysis (SC06)*, 2006.

[11] Z. Liu and D. W. . Cheung, "Oblivious routing for lc permutations on hypercubes," *Parallel Computing*, no. 25, pp. 445–460, 1999.

[12] Varga, András, "The OMNeT++ discrete event simulation system," *Proceedings of the European Simulation Multiconference (ESM'2001)*, 2002.

 ISBN:978-1-61208-113-7

A Review of IPv6 Multihoming Solutions

Habib Naderi
Department of Computer Science
University of Auckland
Auckland, New Zealand
hnad002@aucklanduni.ac.nz

Brian E. Carpenter
Department of Computer Science
University of Auckland
Auckland, New Zealand
brian@cs.auckland.ac.nz

Abstract - Multihoming is simply defined as having connection to the Internet through more than one Internet service provider. Multihoming is a desired functionality with a growing demand because it provides fault tolerance and guarantees a continuous service for users. In the current Internet, which employs IPv4 as the network layer protocol, this functionality is achieved by announcing multihomed node prefixes through its all providers. But this solution, which employs Border Gateway Protocol, is not able to scale properly and adapt to the rapid growth of the Internet. IPv6 offers a larger address space compared to IPv4. Considering rapid growth of the Internet and demand for multihoming, the scalability issues of the current solution will turn into a disaster in the future Internet with IPv6 as the network layer protocol. A wide range of solutions have been proposed for multihoming in IPv6. In this paper, we briefly review active solutions in this area and perform an analysis, from deployability viewpoint, on them.

Keywords - IPv6, Multihoming

I. INTRODUCTION

The rapid growth of the Internet, during recent years, and known limitations in its native network protocol have raised some concerns among experts about the future. IPv4 addresses will run out in the near future. It is a big obstacle to the development of the Internet. One proposed solution, and the most promising one, is replacing IPv4 with a new protocol, which is able to resolve IPv4 issues. Early deployments and experiments have shown that IPv6 is stable and reliable enough to replace IPv4, but a practical and incremental deployment plan and also a reasonable solution for multihoming seem necessary. Multihoming has been an open problem for 35 years since the invention of the Internet [1] and no perfect solution has been proposed for that during these years.

Multihoming is simply defined as having connection to the Internet through more than one Internet Service Provider (ISP). Multihoming can be implemented at host or site level. A host with two or more independent connections to the Internet is called a multihomed host. A multihomed host is able to detect failures and move established communications from the failed path to one of the available working paths. A site with two or more independent connections to the Internet is called a multihomed site. A multihomed site provides multihoming functionality for its hosts. Hosts are usually unaware of the existence of multihoming in this case. Multihoming is a desired functionality because it provides fault tolerance and guarantees a reliable connectivity for users. So many users, all around the world, are interested to use and benefit from this functionality.

To achieve this functionality in the current Internet, Border Gateway Protocol (BGP) features are employed. A multihomed site acquires its Provider Independent (PI) or Provider Aggregatable (PA) prefix and then announces it through all its providers [2]. In case of PI addresses, the site's prefix appears in the Internet core routing system more than once. In other words, Internet core routers have to process more than one entry for this prefix in their routing tables. An observational study in 2009 [3] showed that employing techniques like CIDR, which make address aggregation possible, have been very helpful to keep the growth of BGP4 table size roughly proportional to the square root of the public Internet size during past years. According to another study [4], multihoming and load balancing have been two major sources of fragmentation and deaggregation of BGP4 announcements. A study in 2005 [5] showed that 20% of entries in the global routing table are associated solely with multihoming. So, as the number of multihomed sites grows rapidly, the routing table size will become a serious issue in the future. Although using PA addresses can avoid the routing table explosion problem, hosts need to be multiaddressed, which creates difficult new issues with ingress filtering, renumbering and session survivability [6].

One major concept, which is employed by most proposed solutions is the separation of identity and location. One of the assumptions in traditional IP design was static topology. So, an object's identity and location were combined into a single protocol element called *IP address*. In IP architecture, identity is the address, which also describes the location. But, new studies showed that we need to separate these two roles. *Identity* uniquely identifies a stack within an end-point, where *Location* identifies the current location of the identity element within the network. It makes it possible to define a multihomed end point with one identity and different locators. The upper layers of protocol stack will deal with identity whereas lower layers should struggle with set of locators. In other words, the upper layer does not need to be aware of multihoming and the service would be transparent to it.

 ISBN:978-1-61208-113-7

A wide range of solutions have been proposed for multihoming in IPv6. These solutions can be categorized in five major categories [7]: Routing Approach, Mobility Approach, Identity Protocol Element, Modified Protocol Element and Modified Site Exit and Host Behaviors.

There are also other views for classifying the proposed solutions. They can be classified according to the location of required modification, i.e. hosts, routing system or both, or according to the network protocol stack element, i.e. network, transport or session, which is affected. In host based solutions, multihoming is implemented in hosts and the routing system is unaware of it. All required information is stored and managed by the host. In routing system based solutions, the routing system is responsible for providing multihoming functionality and storing and managing required information. Hosts are unaware of multihoming in this case. In mixed solutions, multihoming functionality is split across hosts and routers and each component should take care of its own functions and information.

The structure of this paper is as follows. Section II presents a brief overview of related works in the area. Section III presents proposed solutions, which are active and have a chance to be selected as the standard solution. Although some solutions discussed in this paper have not been proposed specifically for multihoming, in all of them multihoming is considered as an important feature. Section IV analyzes these active solutions from deployability view point. We conclude our work in section V.

II. RELATED WORK

Pekka Savola et al. presented the result of their survey on site multihoming in IPv6 in [8]. They presented an overview of proposed solutions along with motivations and challenges in this area and tried to show that solutions for IPv4 are not well structured enough to be applied to IPv6. Cedric de Launois et al. [9] surveyed main solutions for IPv6 multihoming, which had been proposed to IETF over the period of 2000-2005. They also compared the solutions and presented their advantages and disadvantages. The results of a comparative analysis, by Shinta Sugimoto and et al., of two host-centric solutions, SHIM6 and SCTP, were presented in [10]. They specifically focused on architecture, failure detection and security. Jun Bi et al. [11] presented a summary of IPv4 multihoming solutions. They also reviewed and analyzed a number of IPv6 site multihoming approaches and chose SHIM6 as the most promising solution. Richard Clayton [12] analyzed multihoming from an economic viewpoint.

III. ACTIVE SOLUTIONS IN THE AREA

Although a wide variety of solutions for IPv6 multihoming have been proposed during past years, there is no agreement in the research and technical community upon choosing one of them as the best solution. Scalability has been the main concern and avoiding huge routing tables has been one of the most important goals in this area. The identifier-locator separation technique is considered as fundamental for this problem and has been employed by a majority of the solutions.

Identifier-locator separation can be implemented in different ways. Deering [13], based on an earlier proposal [14], proposed dividing IP address space into two portions, one portion to be used as the set of end-system identifiers and the other portion as wide-area locators. Hosts put identifiers, as source and destination addresses, in packets, and border routers encapsulate these packets with an outer header, which contains locators. This scheme is generically called *map-n-encap*. Mapping identifiers to locators needs an infrastructure, which needs to be fast and reliable. Map-n-encap technique also increases the size of packets, which may cause packet fragmentation, if it exceeds MTU. Another way to implement identifier-locator separation is cutting the 16-byte IPv6 address in half and then assigning one half to identifier and another half to locator. The locator part can be rewritten by the routing system, while the identifier part is fixed and unique. Hosts ignore the locator part and just use the identifier part. This approach was initially proposed by O'Dell [15] and is referred as "8+8". The positive aspect of both approaches is that the delivered packet would be identical to the sent packet although the header is rewritten by exit routers. It avoids undesirable side-effects, which are caused by similar techniques like Network Address Translation (NAT) in IPv4 [16]. Because of perceived security issues, the 8+8 proposal was not updated, but the idea has been widely used in other proposals.

Other approaches like using geographically based address prefixes [17], transport protocols with multihoming support like Stream Control Transmission Protocol (SCTP) [18] and introducing an additional level of identifier above the IP address, namely HIP [19] have also been proposed. From 2001 to 2003, more than 35 drafts related to IPv6 multihoming were produced in IETF to cover different classes of solutions [20]. After reviewing these proposals, SHIM6 [21] was selected as a standard solution. SHIM6 is a host centric solution, compatible with IPv6 and its routing architecture, which simulates identifier-locator separation. SHIM6 is not an attractive solution for service providers because it does not provide a powerful set of traffic engineering features. Using PI addresses were considered in some solutions when Regional Internet Registries removed restrictions for allocating PI prefixes. Some early IPv6 adopters used IPv4 style solutions, which raised the concern about routing table explosion problem. So, after an Internet Architecture Board workshop and report [22], new technical proposals were produced. Some of them are still active and under development [23]. LISP, ILNP, NAT66, MPTCP, continued work on HIP, name-based transport and SHIM6 are the main proposals, which are summarized and analyzed in this paper.

LISP (Locator/ID Separation Protocol [24]) is a map-n-encap solution, which is with an active IETF Working Group. LISP inserts a new network layer below the host stack network layer. The host network stack works with

 ISBN:978-1-61208-113-7

EIDs (End-point Identifiers) while the new layer works with RLOCs (Routing Locators). EID, which is a non-routable IP address, uniquely identifies a host while RLOCs are routable PA addresses, which should be easily aggregatable in the BGP4 system. LISP has two major components: data plane, which performs map-n-encap operation, and control plane, which is the EID-to-RLOC mapping system. The map-n-encap process is performed by LISP routers, ETR (Egress Tunnel Router) and ITR (Ingress Tunnel Router). ETRs perform decapsulation and ITRs are responsible for encapsulation. A fast and reliable mapping system should provide assistance for ITRs so that they can encapsulate outgoing packets in an outer header, which contains RLOCs. Incremental deployment, which needs interoperability with existing unmapped Internet, is a tricky issue [25]. One proposed solution for this problem is using proxy tunnel routers, which announce a large range of EIDs in an aggregated form. The communication between LISP and non-LISP hosts will then become possible through these proxies.

ILNP (Identifier Locator Network Protocol [26]), a direct descendant of 8+8 [15], is a network protocol, which has been designed based on identifier locator separation approach. To be incrementally deployable, designers propose building that upon IPv6. Packet headers for ILNP and IPv6 are nearly identical but, like 8+8, 64 bits of address is used as locator followed by a 64-bit identifier. The identifier names a node, not an interface, and is in IEEE EUI-64 format and is not used for forwarding. The identifier is not required to be globally unique, but a unique identifier would be very helpful. Hosts should be aware of ILNP to be able to detect failures and recover from them. ICMP protocol is used for locator updates and four new resource records should be supported by DNS.

NAT66 [27] is a stateless version of NAT44 (NAT for IPv4). Like NAT44, the source address is overwritten by NAT66 node before sending a packet out and the destination address is overwritten before sending a received packet in. NAT66 does not include port mapping, as there is an external address for every internal address. Employing NAT66 on the border router of a multihomed site enables address mapping from different external addresses to the same set of internal addresses. Switching between providers is done by changing external address in the NAT66 mapping process. Address mapping is algorithmic and checksum-neutral. Thus there is no need to maintain any per-node or per-connection state; address rewriting keeps the checksum in the transport layer unchanged. Thus there is no need to modify transport layer headers. NAT66 also allows internal nodes to be involved in peer-to-peer communications.

MPTCP (MultiPath TCP [28]) is an extension to traditional TCP, to enable it to use multiple simultaneous paths between multihomed/multiaddressed peers. The aim of MPTCP is to improve resource utilization and failure tolerance. MPTCP is a set of features on top of TCP, meant to be backward compatible, so as to work with middle boxes (e.g. NAT, firewall, proxy) and legacy applications and systems without affecting users. A MPTCP connection is started like a regular TCP connection. Then, if extra paths exist, additional TCP connections (subflows) will be created. MPTCP operates such that all these connections look like a single TCP connection to the application. There are two major differences between MPTCP and transport protocols like SCTP. First, MPTCP preserves the TCP socket interface, so it is fully compatible with existing TCP applications. Second, it uses all available address pairs between communication hosts simultaneously, and spreads the load between working paths using TCP-like mechanisms.

HIP [19] is a host-based solution for secure end-to-end mobility and multihoming, using an identity/locator split approach. In HIP, IP addresses are used as locators but host identifier is the public key component of a private-public key pair. Host identity is a long term identity so it can be used for looking up locators. Host identity is created by the host itself and can be stored in DNS to be searchable by other hosts. Each host has one host identity but can have more than one host identifier. [29] proposes a common socket API extension for HIP and SHIM6 since from upper layer's viewpoint, they look similar.

Name-based transport [30] is an evolution of the existing socket interface, which hides multihoming, mobility and renumbering from applications. Applications do not need to struggle with addresses. They can simply use domain names and leave the management of IP addresses in communication sessions to the operating system.

SHIM6 [21] is a host-centric solution, chosen by IETF as an engineering solution, for IPv6 multihoming. SHIM6 uses identity/locator scheme but does not define a new name space. IPv6 addresses are used as identifier and locator. Initial connection, similar to non-shim6 connections, uses one of the available host's IP addresses. This address will play the role of identifier, which is called ULID (Upper Layer ID), during the communication lifetime. ULID is associated with a list of the host's other IP addresses, referred to as locators. SHIM6 inserts a shim layer on top of the IP routing sub-layer and under IP endpoint sub-layer. This layer performs a mapping between ULID and locator(s). SHIM6 employs a separate protocol, called REAchability Protocol (REAP) [31], for failure detection and recovery. The recovery process is independent from and transparent to upper layer protocols. To benefit from the mentioned functionality, both ends of a communication should implement SHIM6. Also, hosts need to be multiaddressed.

IV. ANALYSIS

Deployability is a key attribute for new Internet protocols. In this section we analyze the active solutions reviewed in section III from a deployability viewpoint. We have considered seven important aspects in our analysis: scalability, amount of required modifications, security, traffic engineering, deployment cost, ease of renumbering and code availability.

 ISBN:978-1-61208-113-7

LISP: RLOCs are assumed to be PA addresses, which are aggregatable in the BGP4 system. So, LISP is considered as a scalable solution. To deploy LISP, no change is required within sites or the Internet core routing system. Modifications are limited to border routers (xTRs). A mapping system, like LISP-ALT [32], is also required to maintain EID to RLOC mappings. Communications between xTRs are protected by using a 32-bit nonce. This technique only provides a basic protection. In fact, LISP and LISP-ALT are not more secure than BGP. The mapping system should support priorities and weights for each locator. Using this information, LISP is able to provide powerful facilities regarding traffic engineering and load sharing. Like other map-n-encap approaches, LISP suffers from encapsulation overhead. Encapsulation increases the packet size and probability of fragmentation, which may have negative impact on performance. IPv6 routers do not perform fragmentation and drop the packets larger than MTU. So, there is a possibility that large LISP encapsulated packets are dropped by IPv6 routers. LISP designers propose some solutions for this problem although, based on informal surveys, they believe that majority of Internet transit paths support a MTU of at least 4470 bytes and there is no need to be worried about this problem. Depending on the mapping system technology, a mapping process may impose an overhead on routing time and traffic. Employing a proper solution for interoperability, as mentioned in section III, LISP can be deployed incrementally. Renumbering only affects xTRs and mapping database. A fast mechanism is required for updating the mapping database, in case of renumbering, to avoid out of date responses to mapping requests. Two implementations are available for LISP: OpenLISP and LISP for IOS (from Cisco).

ILNP: ILNP is mainly implemented in hosts. Hosts can be multiaddressed and by using PA addresses, address aggregation is completely possible. So, ILNP is considered as a scalable solution. To deploy ILNP, hosts should be modified. Also, support for new resource records (I, L, PTRI and PTRL) should be added to DNS. ILNP employs ICMP protocol for locator change notification. Support for a new message called *Locator Update* needs to be added to ICMP. Although ILNP encourage applications to use FQDNs instead of IP addresses, legacy applications would still be able to work with ILNP if required APIs for conversions between FQDN and IP addresses are provided. ILNP employs IPSec to improve the security of communications. It does not include locators in authentication header, so changing locators does not affect the security of communications. To provide proper traffic engineering facilities, ILNP authorizes edge routers to rewrite locators in packet headers and enforce TE policies. ILNP is compatible with pure IPv6 so an approach like dual stack seems possible for incremental deployment. To handle a renumbering, DNS records should be updated because Identifier-locator mappings are stored in DNS. Only a research demonstration implementation of ILNP, from the University of St Andrews, is available at the moment.

NAT66: With NAT66, sites are able to use PI addresses as internal addresses within the site and PA addresses, which are aggregatable in the Internet routing system, as external addresses. Although internal addresses are accessible from outside, but they don't need to appear in the Internet core routing tables, thanks to NAT66 two- way mapping algorithm. So, NAT66 can be considered as a scalable solution. To deploy NAT66, no modification is required in hosts and routers. Just a NAT66 device is required to be installed on the site's exit border. Two-way address mapping enables hosts behind a NAT66 device to be accessed from outside and involved in peer-to-peer communications. It makes NAT66 less secure than NAT44 but the result is not worse than regular IPv6 communications. NAT66 does not offer any specific feature for traffic engineering but NAT66 devices could be improved to enforce TE policies. Address translation imposes a processing overhead on packet forwarding. To use NAT66 address mapping algorithm, both internal and external prefixes should be /48 or shorter to have at least 16 bits available for subnet; otherwise checksum neutrality cannot be guaranteed. Renumbering is easy, only the NAT66 device should be modified to use new prefix(es). NAT66 is not able to preserve established communications in case of renumbering and failure. There is no implementation available for NAT66 at the moment.

MPTCP: MPTCP extends TCP capabilities and allows hosts to benefit from parallel flows to improve the performance and network utilization. MPTCP needs hosts to be multiaddressed and addresses are assumed to be PA addresses to take care of scalability. To deploy MPTCP, only hosts need to be modified. MPTCP is backward compatible with TCP, so TCP applications are able to use it easily without need to any change. MPTCP designers have tried to keep MPCTP as secure as TCP but multipath feature has opened some security concerns [33]. MPTCP allows hosts to enforce their preferences for spreading their traffic over different paths, but there is no way for receivers to change these preferences. Some solutions like ECN and fake congestion signals [34] have been proposed for this problem. MPTCP is backward compatible with traditional TCP, so incremental deployment is possible. But to benefit from multipath features, both end of communication should support MPTCP. One of the host's IP addresses, which is used for establishing connection, plays the role of identifier and also locator for one of subflows. In case of renumbering, such subflows can cause confusion and security problems. Two versions of MPTCP, based on LinShim6 code base, have been implemented in Université Catholique de Louvain. Both are still incomplete.

HIP: HIP allows hosts to be multiaddressed and addresses are assumed to be PA addresses. So, address aggregation is possible without any change in routing system which makes HIP a scalable solution. HIP is a host-centric, solution and major modifications should be implemented in hosts. To maintain host identifiers, DNS or a PKI (Public Key Infrastructure) is required. To benefit from HIP features, applications should use an extended socket interface, which has been proposed for this purpose [29]. Another version of

 ISBN:978-1-61208-113-7

HIP, opportunistic HIP, has been proposed for situations where DNS or PKI is not available. HIP employs IPSec to provide a secure media for communications. There is no specific facility for traffic engineering in HIP. There are some issues regarding incremental deployment of HIP [35]. HIP is able to change locators without breaking communications. To handle renumbering, host identifiers should be updated in DNS/PKI. There are five implementations for HIP: OpenHIP, HIP for Linux, HIP for inter.net, InfraHIP and pyHIP.

Name-based Sockets: Name-based sockets implement an identifier-locator separation scheme by allowing applications to use domain names instead of IP addresses. IP addresses are assumed to be managed by the operating system. Using PA addresses, address aggregation is easily possible so, this solution can be considered as a scalable solution. To deploy Name-based sockets, hosts networking stack should be modified to support required features. No modification is required in routing system. Name-based sockets are vulnerable to domain name spoofing, redirection and flooding attacks. Solutions like using additional forward lookups in DNS for verifying domain names and exchanging random numbers, in case of redirection, have been proposed to protect Name-based sockets against mentioned attacks. Name-based sockets are not intended to improve IPv6 security; they just try to keep the level of security at the same level as today's Internet. This solution does not provide any specific facility for traffic engineering. Name-based sockets are backward compatible to traditional socket interface, so incremental deployment is possible. Name-based sockets provide required mechanisms for changing locators without breaking communication sessions. So, renumbering is easy and just needs an update to DNS. A prototype of name-based sockets has been implemented as a result of collaboration between Ericsson, Tsinghua University and Swedish Institute of Computer Science.

SHIM6: SHIM6 is a host-centric solution, which is able to provide multihoming functionality for multiaddressed hosts. If addresses are PA addresses, address aggregation would easily be possible. So, SHIM6 is considered as a scalable solution. SHIM6 is implemented in hosts and doesn't need any change in the routing system. SHIM6 is not intended to improve the security of the IPv6 communications. HBA/CGA, context tag and a 4-way handshake mechanism for context establishment have been employed to help SHIM6 not to downgrade the security. SHIM6 provides some simple mechanisms regarding traffic engineering. Hosts are able to notify the other end of communication about their preferences among available locators. It is a host level mechanism and site administrators need other mechanisms for enforcing traffic engineering policies in their sites. [36] proposes some improvements to SHIM6 for enhancing its traffic engineering capabilities. A SHIM6 capable host is able to communicate with non-SHIM6 hosts. Thus, incremental deployment is possible, although SHIM6 is unable to activate its capabilities in these cases. SHIM6 is able to handle locator changes on the fly, so handling renumbering is easy. If a renumbered prefix is in use, the corresponding context can still continue its work. But, such contexts are a source of confusion and security issues. Two implementations are available for SHIM6: LinShim6 and OpenHIP.

Figure 1 shows a table summarizing characteristics of the described solutions. Our analysis can be summarized as follows: SHIM6, HIP, MPTCP, ILNP and name-based sockets are, in fact, solutions for host multihoming while LISP and NAT66 are considered as site multihoming solutions. The amount of required modifications for deploying a solution is an important factor. Solutions, which need fewer modifications would be more desirable since they offer less deployment cost. LISP offers some precise features for traffic engineering, other solutions just propose some general guidelines and possibilities. Traffic Engineering is an important feature from administrator's viewpoint as it enables them to control site's incoming and outgoing traffic. LISP and HIP have some issues with incremental deployment. As the Internet is a widespread network, incrementally deployable solutions have a higher chance to be adopted. Only NAT66 is not able to preserve communications in case of failure and renumbering, although SHIM6 and MPTCP also have some issues with renumbering in special cases. Solutions, which make renumbering simple are more desirable from a site administrator's viewpoint because they offer more flexibility in changing service providers. From a technical viewpoint, it seems that ILNP and LISP offer a more complete set of features compare to other solutions. The co-chairs of the IRTF RRG have recommended the work on ILNP be pursued toward a routing architecture in which multihoming will be one of the main features [23].

V. CONCLUSION

This paper presents a review of active multihoming solutions for IPv6. Although a large number of solutions have been proposed for this problem, few of them satisfy necessary technical requirements and therefore have a chance to be chosen, by the technical community, as the standard solution. We summarized and analyzed seven important solutions, which are active in this area. Results of our analysis show that each solution has its own drawbacks and weak points so that it is difficult to choose one of them as "the perfect solution". On the other hand, some characteristics, which are positive from technical viewpoint, do not seem to be easily deployable in the Internet. For example, considering number of modifications as a deployability parameter, host-based solutions need modifications only in one component: hosts. Technically, it might be possible to consider this class of solutions as "simply deployable" but such changes cannot be made without close cooperation of OS and networking software vendors. Also, end users should be convinced to pay the cost of such updates to their hosts. It seems that more research and effort is still needed for achieving a scalable, deployable, manageable and secure solution for IPv6 multihoming.

 ISBN:978-1-61208-113-7

Solution / Characteristic	LISP	ILNP	NAT66	MPTCP	HIP	NBS	SHIM6
Product Modifications	ER	H, SP, A*, ER*	None	H	H, A, SP*	H, A	H, A*
Security(Compare to BGP4)	Similar	Stronger	Similar	Similar	Stronger	Similar	Similar
TE (Compare to BGP4)	Stronger	Similar	Weaker	Weaker	Weaker	Weaker	Weaker
Incremental Deployment	Possible with Conditions	Possible	Possible	Possible	Possible with Conditions	Possible	possible
Renumbering without breaking established communications	Possible	Possible	Impossible	Possible with Conditions	Possible	Possible	Possible with Conditions
New Component	Mapping System	None	NAT Device	None	PKI*	None	None

*: optional A: Application ER: Edge Router H:Host SP: Services and Protocols

Figure 1. Summary of characteristics of the discussed solutions

REFERENCES

[1] L. Pouzin, Interconnection of packet switching networks, 7th Hawaii International Conference on System Sciences, Supplement, 1974.

[2] J. Abley, K. Lindqvist, E. Davies, B. Black, and V. Gill, IPv4 Multihoming Practices and Limitations, Internet RFC 4116, July 2005.

[3] B. E. Carpenter, Observed Relationships between Size Measures of the Internet, ACM SIGCOMM CCR, 39(2) (April 2009).

[4] T. Bu, L. Gao, and D. Towsley. On characterizing BGP routing table growth. Computer Networks, 45(1):45–54, 2004.

[5] X. Meng, Z. Xu, B. Zhang, G. Huston, S. Lu, and L. Zhang, IPv4 Address Allocation and the BGP Routing Table Evolution, ACM SIGCOMM Computer Communication Review, 35(1), 2005.

[6] J. Abley, B. Black, and V. Gill, Goals for IPv6 Site-Multihoming Architectures, Internet RFC 3582, August 2003.

[7] G. Huston, Architectural Approaches to Multi-homing for IPv6, Internet RFC 4177, September 2005.

[8] P. Savola and T. Chown, A Survey of IPv6 Site Multihoming Proposals, 8th International Conference on Telecommunications (conTEL), 2005.

[9] C. de Launois and M. Bagnulo, The Paths Toward IPv6 Multihoming, IEEE Communications Survey 8 (2006) 38-50.

[10] S. Sugimoto, R. Kato, and T. Oda, A Comparative Analysis of Multihoming Solutions, IPSJ SIG Technical Report, 2006.

[11] J. Bi, P. Hu, and L. Xie, Site Multihoming: Practices, Mechanisms and Perspective, Future Generation Communication and Networking (FGCN) 1 (2007) 535-540.

[12] R. Clayton, Internet Multi-Homing Problems: Explanations from Economics, Eighth Annual Workshop on Economics and Information Security (WEIS09), London, UK, June 24-25, 2009.

[13] S. Deering, The Map & Encap Scheme for scalable IPv4 routing with portable site prefixes, presentation at IETF35, Los Angeles, March 4-8, 1996.

[14] R. Hinden, New Scheme for Internet Routing and Addressing (ENCAPS) for IPNG, Internet RFC 1955, June 1996.

[15] M. O'Dell, 8+8 - An Alternate Addressing Architecture for IPv6, Internet Draft (work in progress), 1996.

[16] K. Egevang and P. Francis, The IP Network Address Translator (NAT), Internet RFC 1631, May 1994.

[17] F. Baker, A Business Model For Metro Addressing, Internet Draft (work in progress), 2001.

[18] R. Stewart, Stream Control Transmission Protocol, Internet RFC 4960, September 2007.

[19] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson, Host Identity Protocol, Internet RFC 5201, April 2008.

[20] List of Internet-Drafts relevant to the Multi6-WG, http://ops.ietf.org/multi6/draft-list.html (last visited 2010-02-24)

[21] E. Nordmark and M. Bagnulo, Shim6: Level 3 Multihoming Shim Protocol for IPv6, Internet RFC 5533, June 2009.

[22] D. Meyer, L. Zhang, and K. Fall, Report from the IAB Workshop on Routing and Addressing, Internet RFC 4984, September 2007.

[23] Li, T. (ed.), Recommendation for a Routing Architecture, Internet Draft (work in progress), 2010.

[24] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, Locator/ID Separation Protocol (LISP), Internet Draft (work in progress), 2010.

[25] D. Lewis, D. Meyer, D. Farinacci, and V. Fuller, Interworking LISP with IPv4 and IPv6, Internet Draft (work in progress), 2009.

[26] R. Atkinson, ILNP Concept of Operations, Internet Draft (work in progress), 2008.

[27] M. Wasserman and F. Baker, IPv6-to-IPv6 Network Address Translation (NAT66), Internet Draft (work in progress), 2010.

[28] A. Ford, C. Raiciu, and M. Handley, TCP Extensions for Multipath Operation with Multiple Addresses, Internet Draft (work in progress), 2009.

[29] M. Komu, M. Bagnulo, K. Slavov, and S. Sugimoto, Socket Application Program Interface (API) for Multihoming Shim, Internet Draft (work in progress), 2009.

[30] J. Ubillos, M. Xu, Z. Ming, and C. Vogt, Name-Based Sockets Architecture, Internet Draft(work in progress), 2010.

[31] J. Arrko and I. Van Beijnum, Failure Detection and Locator Pair Exploration Protocol for IPv6 Multihoming, Internet RFC 5534, June 2009.

[32] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis, LISP Alternative Topology (LISP+ALT), Internet Draft(work in progress), 2010.

[33] M. Bagnulo, Threat analysis for Multi-addresses/Multi-path TCP, Internet Draft (work in progress), 2009.

[34] K. Ramakrishnan, S. Floyd, and D. Black, The Addition of Explicit Congestion Notification (ECN) to IP, Internet RFC 3168, September 2001.

[35] T. Henderson, P. Nikander, and M. Komu, Using the Host Identity Protocol with Legacy Applications. Internet RFC 5338, September 2008.

[36] M. Bagnulo, A. Garcia-Martinez, and A. Azcorra, BGP-like TE Capabilities for SHIM6, Proceedings of the 32nd EUROMICRO-SEAA'06, 2006.

 ISBN:978-1-61208-113-7

An Analytic and Experimental Study on the Impact of Jitter Playout Buffer on the E-model in VoIP Quality Measurement

Olusegun Obafemi
School of Information Technology
Illinois State University
Normal IL 61790 USA
oeobafe@ilstu.edu

Tibor Gyires
School of Information Technology
Illinois State University
Normal IL 61790 USA
tbgyires@ilstu.edu

Yongning Tang
School of Information Technology
Illinois State University
Normal IL 61790 USA
ytang@ilstu.edu

***Abstract*—Over the years, the quality in Voice over IP (VoIP) applications has been defined from the perspective of VoIP service providers but not necessarily from the perspective of end-users. Conversational quality as perceived by end-users is affected by a wide variety of factors that are not exclusive to network performance. While service providers focus their efforts on Quality of Service (QoS) in VoIP, end-users expect improvement in Quality of Experience (QoE). The measurement of QoE is significantly challenging due to the diverse nature of factors that collectively determine QoE. It is highly desired to have a more comprehensive and accurate quality measurement model that can provide objective measurement of QoS in real-time while simultaneously offering QoE measurement as well. Thus, unveiling the relevance of various factors in a quality measurement model is crucial. In this paper, we study a commonly used ITU-T recommended quality measurement model: the E-model with a focus on how an important but ignored parameter jitter playout buffer affects the measurement accuracy by conducting an analytical and experimental approach. Our study shows that adaptive jitter playout buffering does affect the measurement of end user perceived VoIP quality. Thus, the E-Model that excludes the impact of the dynamic jitter playout buffer cannot provide accurate measurement on how an user perceives conversational quality in VoIP application.**

Keywords- Quality of Service, Quality of Experience, Jitter Buffer, Adaptive Playout Buffering, Conversational Quality, VoIP.

I. INTRODUCTION

Quality of Service (QoS) in VoIP prioritizes voice traffic at the expense of traffic from other packet types. The factors of QoS mainly include jitter, packet loss and latency. Jitter reflects the delay variation that affects voice IP packets as they travel through a network. Packet loss occurs when voice packets are dropped due to various reasons (e.g., congestions) in the network. Latency or delay is due to several factors, including physical distance between a sender and a receiver, the number of router hops the packets have to pass through, and the packet processing time in the network.

The QoS provisioned by VoIP service providers is not necessarily proportional to the Quality of Experience (QoE) perceived by end users [4]. VoIP QoE describes the satisfaction of end users with respect to conversations using VoIP technology. The tremendous increase of VoIP users in the recent years makes the satisfaction of users an equally important metric, compared to the QoS related network performance metrics.

The measurement of QoE is significantly challenging due to the diverse nature of factors that collectively determine QoE. Several ITU recommended methods are commonly used, namely Mean Opinion Score (MOS) [4], Perceptual Evaluation of Speech Quality (PESQ) [5] and E-Model [9], which we will discuss later. However, the MOS based user satisfaction measurement has several inherent shortcomings [1], including the unrepeatable nature of the tests and the fact that the tests are not scalable. Similarly, the PESQ and E-Model cannot provide accurate objective measurement on user perceived quality as well as user satisfaction [8]. PESQ does not consider several factors such as end-to-end delay in its computations. On the other hand, the E-Model, although it includes conversational voice quality measurements perceived by end-users, it still ignores the dynamic nature of IP networks and subsequently excludes the impairment caused by inappropriate playout buffering.

It is highly desired to have a more comprehensive and accurate quality measurement model that can provide objective measurement of QoS in real-time while simultaneously offering QoE measurement as well. Thus, unveiling the relevance of various factors in a quality measurement model is crucial.

In this paper, we study a commonly used ITU-T recommended quality measurement model: the E-model with a focus on how an important but ignored parameter, the jitter playout buffer, affects the measurement accuracy by conducting an analytical and experimental approach. Our study shows that adaptive playout buffering does affect the measurement of end user perceived VoIP quality. Thus, the E-Model that excludes the impact of the dynamic playout buffer cannot provide accurate measurement on end user perceived conversational quality in VoIP applications.

The rest of the paper is organized as follows. Section II first presents a comprehensive literature review. Then, Section III discusses the three ITU-T recommended models that can provide subjective and objective measurements of

 ISBN:978-1-61208-113-7

conversational quality. Section IV analyzes several objective measurement factors that determine quality of service. Section V provides the experimental study via simulations. Section VI concludes our study and briefly discusses our future work.

II. Related Work

There are numerous approaches proposed to objectively measure speech quality in VoIP. Robinson and Yedwab [10] proposed a Voice Performance Management system to monitor call quality in real-time by proactively monitoring, alerting, troubleshooting and reporting network performance problems. Robinson and Yedwab [10] concluded that only packet loss, jitter and latency show the correlations between QoS and QoE.

Gierlich and Kettler [11] provided insight into the impact of different network conditions and the acoustical environment on speech quality. Testing techniques for evaluating speech quality under different conversational aspects were also described. Gierlich and Kettler [11] argued that there is no single number that can objectively indicate speech quality; and pointed out that overall speech quality is a combination of different single values from different speech quality parameters. Wang et. al., [12] designed and implemented a QoS-provisioning system that can be seamlessly integrated into current Cisco VoIP systems. Wang et. al., [12] also described Call Admission Control (CAC) mechanisms (Site-Utilization-based CAC and Link-Utilization-based CAC) to prevent packet loss and over-queuing in VoIP systems.

Myakotnykh and Thompson [13] described an algorithm for adaptive speech quality management in VoIP communications, which can show a real-time change in speech encoding parameters by varying voice packet sizes or compression (encoding) schemes. The algorithm involves the receiver making control decisions based on computational instantaneous quality level (which is calculated per talkspurt using the E-Model) and perceptual metric (which estimates the integral speech quality based on latency, packet loss and the position of quality degradation period in the call). Myakotnykh and Thompson [13] calculated the maximum achievable quality level for a given codec under specific network conditions, packet playout time, packet delay before jitter buffer and degradation in quality caused by traffic burstiness and high network utilization. The algorithm however results in an increase in average quality without increasing individual call quality.

Raja, Azad and Flanagan [14] designed generalized models to predict degradation in speech quality with high accuracy, in which genetic programming is used to perform symbolic regressions to determine Narrow-Band (NB) and Wide-Band (WB) equipment impairment factors for a mixed NB/WB context. Zha and Chan [15] described two algorithms for objective measurement of speech quality: single-ended (needing only to input the degraded speech signal) and double-ended (needing both the original and degraded speech signals). The algorithm developed by Zha and Chan [15] can objectively measure in real-time speech quality using statistical data mining methods.

Several algorithms have also been proposed to optimize some of the existing ITU-T models. The goal of optimization is to enhance existing models by correcting weaknesses that are identified in the models. Gardner, Frost and Petr [16] proposed an algorithm to optimize the E-Model by considering coder selection, packet loss, and link utilization. The authors however stated that the algorithm would have to be enhanced if used in a wide area network involving multiple users. Mazurczyk and Kotulski [17] proposed an audio watermarking method based on the E-Model and the MOS, which provides speech quality control by adjusting speech codec configuration, playout buffer size and amount of Forward Error Correction (FEC) mechanism in VoIP under varying network conditions.

One of the limitations of the E-model is the fact that the model does not consider the dynamic nature of underlying networks that support VoIP. This limitation is addressed by several authors designing adaptive playout buffering to improve voice quality in VoIP. Most of these studies either optimize the E-Model, the PESQ or combine the PESQ and the E-Model to propose a more holistic solution. Mazurczyk and Kotulski [17] highlighted two problems that are associated with adaptive playout buffering: how to estimate current network status and how to transfer network status data to the sending or receiving side. Wu et. al., [18] admitted that VoIP playout buffer size has long been a challenging optimization problem, as buffer size must balance the dynamics of conversational interactivity and VoIP speech quality. Wu et. al., [18] stated that the optimal playout buffer size yields the highest satisfaction in a VoIP call. Wu et. al., [18] investigated the playout buffering dimensions in Skype, Google Talk and MSN Messenger. Wu et. al., [18] concluded that MSN Messenger produces the best performance in terms of adaptive playout buffering, while Skype does not adjust its playout buffering at all. Narbutt and Davis [19] stated that the management of playout buffering is not regulated by any standard and is therefore vendor specific. Narbutt and Davis [19] proposed a scheme that extends the E-Model and provides a direct link to perceived speech quality. Narbutt and Davis [19] evaluated various playout algorithms in order to estimate user satisfaction from time varying transmission impairments including delay, echo, packet loss and encoding scheme.

III. The Analysis of Quality of Experience Measurement Models

In this section we briefly review three commonly adopted ITU-T recommended QoE measurement methods: Mean Opinion Score (MOS), Perceptual Evaluation of Speech Quality (PESQ) and E-Model.

 ISBN:978-1-61208-113-7

A. Mean Opinion Score

Mean Opinion Score or MOS has been endorsed by ITU-T as a subjective method to evaluate voice transmission quality. The MOS test involves using a group of testers (listeners) to assign a rating to a voice call. The quality is rated on a scale of 1 to 5, with 1 = bad, 2 = poor, 3 = fair, 4 = good and 5 = excellent [2]. The arithmetic mean of the scores provided by all listeners becomes the final MOS value of the voice call. Assessment ratings can also be obtained by clustering the test results as "Good or Better" or as "Poor or Worse", and further calculating the relative ratio or percentage of each type of results. For a given voice call, these results are expressed as "Percentage Good or Better" (%GoB) and "Percentage Poor or Worse" (%PoW) [3]. Table I shows the MOS rating, %GoB, %PoW and the correlation between each rating [4].

Table I: Subjective Ratings for Measuring QoE

User Satisfaction	**MOS** (5)	**%GoB** (100)	**%PoW** (0)
Very Satisfied	4.3-4.4	97.0-98.4	0.2-0.1
Satisfied	4.0-4.29	89.5-96.9	1.4-0.19
Some Dissatisfied	3.6-3.9	73.6-89.5	5.9-1.39
Many Dissatisfied	3.1-3.59	50.1-73.59	17.4-5.89
Nearly All Dissatisfied	2.6-3.09	26.59-50.1	37.7-17.39
Not Recommended	1.0-2.59	0-26.59	99.8-37.69

The advantage of the MOS is that it can provide an offline analysis of end-user opinions. However, MOS tests cannot provide an absolute reference for the evaluations; that is, MOS ratings are dependent on the expertise of listeners [1]. Moreover, MOS tests cannot be used in large scale experiments that involve a large number of users because of the involved overhead (e.g., test setup). Moreover, MOS tests are unrepeatable by nature.

B. Perceptual Evaluation of Speech Quality

ITU-T P.862 (PESQ) involves the comparison between the original or reference speech samples and the ones traversed through a test network channel. The more similar the output signal is to the reference signal, the higher the score assigned to the quality of the transmission channel. The comparison result, referred to as the PESQ, is in the range of -0.5 to 4.5 which can be linearly projected to the corresponding MOS score. The PESQ, however, cannot comprehensively represent conversational voice quality due to its exclusion of several network and system parameters including end-to-end delay, echo, listening level, sidetone, loudness loss, Enhanced Variable Rate Codec (EVRC) [5].

C. E-Model

The E-Model is designed to measure the instant user perceived quality instead of the cumulative effect during an entire conversation. The E-Model assumes that individual impairment factors are additive on a psychological scale and combines the cumulative effects of these factors into the Transmission Rating Factor, R, which can be transformed into other quality measures like the MOS, Percentage Good or Better (%GoB) or the Percentage Poor or Worse (%PoW). The R-rating is on a scale of 0 to 100, with high values of R between 90 and 100 interpreted as excellent quality, while lower values of R indicate a lower quality. Values of R below 50 are considered unacceptable and values above 94.15 are assumed to be unobtainable in narrowband telephony. The E-Model measures individual impairment factors at different points in time to compute the R-rating. The value of the R-rating is consequently associated with measurements taken at a given time point and does not reflect the dynamic nature of quality during the entire length of a conversation. The following formula shows the computation of the R-rating:

$$R = R_0 - I_s - I_d - I_e + A \quad (1)$$

R_O represents the basic signal-to-noise ratio, including noise sources such as circuit noise and room noise. The factor I_s is a combination of all impairments which occur simultaneously with the voice signal. The factor I_d represents the impairments caused by delay, and the effective equipment impairment factor I_e represents impairments caused by low bit-rate codecs and packet-losses of random distribution. The advantage factor A corresponds to the user allowance due to the convenience when using a given technology.

The E-Model not only takes in account the transmission statistics (transport delay and network packet loss), but it also considers the voice application characteristics, like the codec quality, codec robustness against packet loss and the late packets discard. However, the impairment due to playout buffer size is simply excluded, which consequently results to the ignorance on how the dynamic varying of the playout buffer size affects QoE throughout an entire conversation. In this paper, we conduct various simulations and experiments to demonstrate that the exclusion of the affect of the playout buffer in the E-model lessen the accuracy of measurements of the conversational quality as perceived by end-users.

IV. The Analysis of Objective Quality Measurement Factors

Quality of Service is determined primarily by latency, packet loss and jitter presented in transmission networks which eventually impose their impacts on user perceived QoE. The ITU-T recommends that as long as the value of latency in a network running VoIP is lower than 150ms, users are expected to be highly satisfied if other factors that may affect network performance are negligible. When latency is between 150ms and 200ms, perceivable performance degradation in quality is expected and becomes more severe with the increase of latency.

VoIP is also sensitive to packet loss, which may result in unintelligible conversation if the packet loss rate is high.

 ISBN:978-1-61208-113-7

The ITU-T recommends that as long as the packet loss rate is less than 1%, users are expected to be highly satisfied if other factors that may affect network performance are negligible. Our simulation via OPNET also shows the same observation. We discuss and show the experimental setup later in Section V. As shown in Figure 1, with increasing packet loss rate, the end user perception of quality gradually decreased.

While gradually increasing the value of packet loss in the network beyond the recommended 1% level, the researcher assigned a score based on the MOS scale. All other impairments were kept at levels where there effects will not be obvious so as not to interfere with the effect of packet loss. Figure 1 shows that as packet loss increased in the transmission network, the end user perception of quality gradually decreased. It is expected that if packet loss decrease in the transmission network and other factors are kept constant, the end user perception of quality will improve. Figure 1 also shows the packet loss in this OPNET simulation was not evident for all three IP phones until after 25 sec.

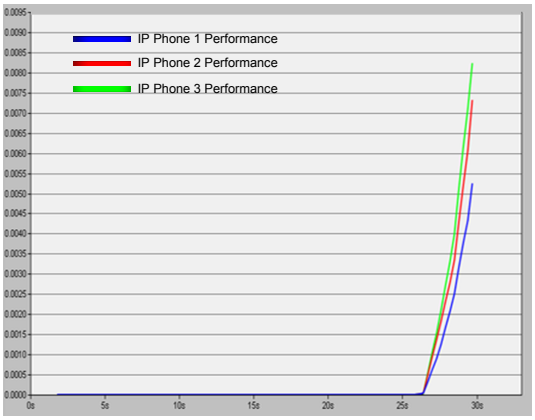


Figure 1: The Effect of Packet Loss Rate on QoE.

V. Experimental Study

This section discusses some of the results that were obtained during experiments that we conducted.

A. Experiment Setup

Figure 2 shows the experimental setup in OPNET that was used to simulate 200 audio conference calls, each with 30s duration. The simulation model included three different participants in the conference. Each of the IP phone nodes had DHCP assigned IP addresses on the same network segment with the two client computer nodes. The experiment simulated a typical packet switched network scenario and used a fixed jitter buffer size that does not dynamically adjust with varying jitter in the network. MOS scores from users associated with the three different IP phone nodes were also simulated and recorded in the experiment.

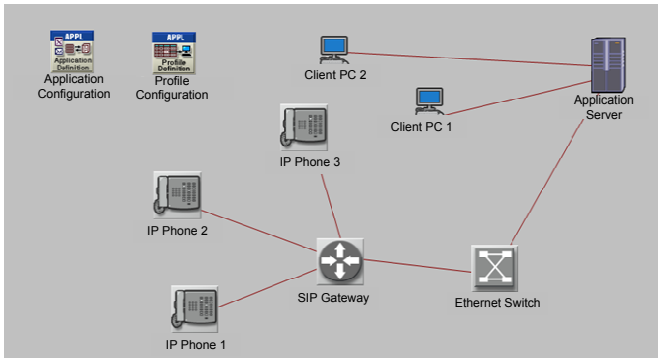


Figure 2: Experiments Setup.

B. Experiment Results

Figure 3 shows the main observations that were recorded during the simulations. The results include MOS values associated with jitter delay, jitter loss and end-to-end delay for the 200 voice calls measured at each IP Phone.

Figure 3(a) shows how static jitter buffer sizes affect the MOS of users. The MOS is taken at every 0.2s during the simulation. The result shows that users' opinion fluctuated throughout the duration of the conversation as a result of the inability of the jitter buffer size to adapt to varying values of jitter in the network.

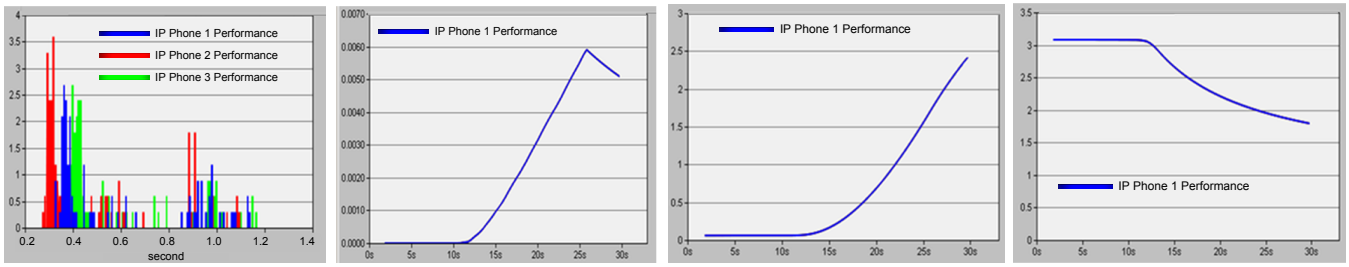
Since the similar performance results were observed for all three IP phones in our experiments, we only show the experiments result from IP phone 1. Figure 3(b) shows the jitter measured at IP phone 1 during the simulation, which demonstrates that after 10s jitter was continuously increasing for about 15s. In the experiment, the jitter varying greatly with the change of various conditions (e.g., available bandwidth, packet loss rate). Figure 3(c) shows the end-to-end delay presented in the network was negligible at the beginning while the network condition was good; however, started increasing significantly after about 15s while the network condition getting worse. Correspondingly, Figure 3(d) shows the MOS varying and reflecting the change of the network condition.

C. Playout Buffering

Mitigating the impact of jitter involves collecting packets in a jitter buffer and playing the packets out relative to the size of the jitter buffer. The size of the jitter buffer affects the end-to-end delay on the network and also the packet loss rate. The size of the jitter buffer may either be kept fixed (static playout buffering) or dynamically adjusted (adaptive playout buffering) with respect to the variations in jitter presented in the network.

Figure 4 shows two different Wireshark interfaces for the same VoIP session. Wireshark allows the playback of different segments of the entire conversation stream. Playing back voice segments and matching the values of jitter present

 ISBN:978-1-61208-113-7



(a) MOS jitter Delay (b) Average Jitter (sec) (c) Average Delay (sec) (d) Average MOS

Figure 3: Voice Application Performance

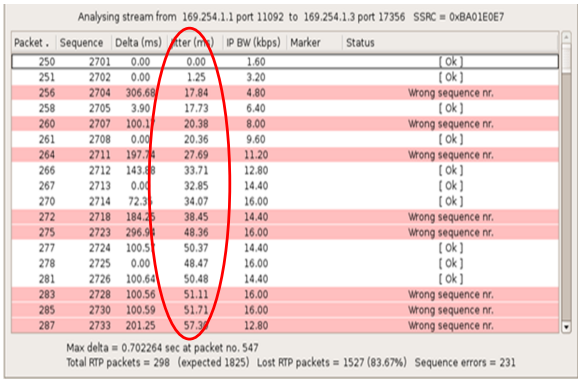


Figure 4: The Wireshark Screenshot Showing the Presence of Jitter.

in each segment confirmed that jitter affects the end user perception of quality.

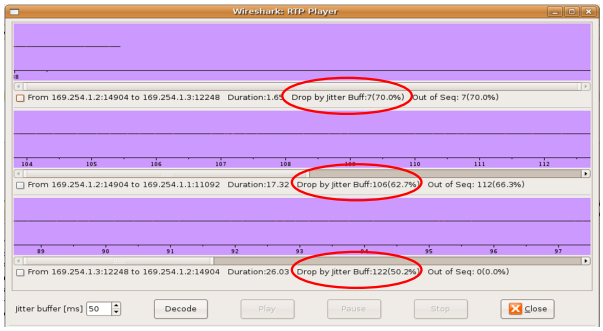


Figure 5: The Wireshark Screenshot Showing the packets dropped by jitter buffer.

Figure 5 shows the results of a typical conversation analyzed using the experimental design shown in Figure 2. Different speech segments of the VoIP session, the source and destination address of the speech segment, together with the number of packets that are out of sequence and those dropped by jitter buffer are shown in Figure 5. Several VoIP sessions were tested using different values of jitter buffer, packet loss, available bandwidth, QoS mechanism, noise reduction, echo cancellation, and different choice of codec.

Static playout buffer size is easier to implement when compared to adaptive playout buffering. However, static playout buffering might sometimes be either too small or too large to handle the fluctuating values of jitter in the network. When the buffer size is too small, slower packets are not played out and dropped. When the buffer size is too large, packets are held longer than necessary and consequently reduce conversational quality.

In contrast to using fixed buffer size, adaptive playout buffering continuously estimates the instant jitter in a network, and further dynamically adjusts the playout buffer to match the varying jitter during a whole conversation. This ensures that packets associated with each conversation segment are played out with the same playout delay. In order to ensure the optimal QoE for end-users, adaptive playout buffering will guarantee that conversational quality is maintained at consistent levels, compared to fluctuating levels associated with static playout buffering.

The output of the E-Model is expected to be a reflection of the quality associated with the entire length of a conversation. However, the computation uses snapshots of impairment values taken at different points within an entire conversation and does not include the affect of the playout buffer. Figure 3(a) shows that the MOS measured at different points during the conversation varied from the beginning of the conversations to the end of the conversations. The measured MOS shown in Figure 3(d) is corresponding to the jitter variations shown in Figure 3(b). At the beginning of the conversations when jitter was zero or almost zero, the highest MOS value was recorded. As jitter increased in the network, the MOS value recorded decreased. Towards the end of the simulation, the MOS value recorded started to increase again.

Adaptive playout buffering would ensure that the variations in jitter in the network are matched by a corresponding value of playout buffer size. This would also ensure that the end-to-end delay in the network is kept at optimal levels and rate of packet loss in the network is kept below the recommended 1%. We will report on the inclusion of the playout buffer in the E-Model in a subsequent paper.

The results of our simulations show that the relevance

 ISBN:978-1-61208-113-7

of adaptive playout buffering should not be ignored in the evaluation of end user perception of conversational quality. Including measurements of adaptive playout buffering into the computations of the E-Model will either reduce the value of the R-rating or keep the value the same, depending on calculated effects of jitter buffer. A modification to the value of the R-rating will provide a more accurate mapping to the MOS. This will in turn give a more accurate prediction of the end user perception of conversational quality.

VI. Conclusion

Through the series of experiments conducted in this study and data analyzed, it was obvious that the quality of experience of end users is a subject that should never be overlooked in the design and development of VoIP solutions. Quality of experience ultimately determines the migration of users from circuit-switched networks to VoIP and the next generation of services. Quality of service related to network configuration and performance was shown in this study as ineffective in ensuring the desired user experience, if other factors that contribute to the quality of experience of end users are not brought to cognizance.

In our future work, we will pursue a longitudinal study tracking the effects of cost of VoIP service, security provision, usability, human behavior, call session and reliability on quality of experience. The goal would be to determine if there are long term effects of these factors on quality of experience. Development of an approach to capture the opinions of end-users about these factors in different parts of the world might be informative. Developing a methodology to conduct subjective tests to indentify how human behavior varies with time during a conversation depending on the quality impairments present would be a future goal.

References

[1] Sat, B., and Wah, B. W. (2009). Analyzing Voice Quality in Popular VoIP Applications. IEEE MultiMedia, vol. 16, no. 1, pp. 46-59. doi:10.1109/MMUL.2009.2.

[2] Zwar, E. J., and Munch, B. (2006). Voice Quality and Network Capacity Planning for VoIP. Retrieved from http://my.gartner.com/portal/server.pt?open=512&objID=260&mode=2&PageID=3460702&resId=493265&ref=QuickSearch&sthkw=G00140936. Last accessed: 5/6/2010.

[3] Narbutt, M., and Davis, M. (2005). Assessing the Quality of VoIP Transmission Affected by Playout Buffer Scheme. 4th International Conference on Measurement of Speech and Audio Quality, Prague, Czech Republic.

[4] International Telecommunications Union (1996). ITU-T P.800. Methods for Subjective Determination of Transmission Quality. Retrieved from http://www.itu.int/rec/T-REC-P.800-199608-I/en. Last accessed: 5/6/2010.

[5] International Telecommunications Union (2007). ITU-T P.862 Corrigendum. Retrieved from http://www.itu.int/rec/T-REC-P.862-200710-I!Cor1/en. Last accessed: 5/6/2010.

[6] Telecommunications Industry Association (2005). Telecommunications, IP Telephony Equipment and Voice Quality Recommendations for IP Telephony. Retrieved from http://ftp.tiaonline.org/TR-41/tr4112inactive/Public/Latest_Revision_of_PN-4689/PN4689LB.pdf. Last accessed: 5/6/2010.

[7] Morris, M. G., Venkatesh, V., Davis, G. B., and Davis, F.D. (2003). User Acceptance of Information Technology: Toward a Unified View. Retrieved from http://proquest.umi.com/pqdlink?vinst=PROD&fmt=6&startpage=-1&ver=1&vname=PQD&RQT=309&did=420086661&exp=10-30-2014&scaling=FULL&vtype=PQD&rqt=309&TS=1257038425&clientId=43838&cfc=1. Last accessed: 5/6/2010.

[8] Becvar, Z., Mach. P., and Bestak, R. (2009). Impact of Handover on VoIP Speech Quality in WiMAX Networks. Eighth International Conference on Networks, icn, pp.281-286, Gosier, Guadeloupe, France.

[9] International Telecommunications Union. (2008). The E-Model. Retrieved from http://www.itu.int/ITU-T/studygroups/com12/emodelv1/tut.htm. Last accessed: 5/6/2010.

[10] Robinson, P. and Yedwab, D. (2009). Voice and Video Application Performance Management in UC Deployments. Retrieved from http://www.ucstrategies.com/uploadedFiles/UC_Information/White_Papers/Psytechnics/Psytechnics UCS_Final_042109%5B1%5D.pdf. Last accessed: 5/6/2010.

[11] Gierlich, H. W. and Kettler, F. (2006). Advanced speech quality testing of modern telecommunication equipment: an overview. Signal Processing, 86(6), 1327 - 1340.

[12] Wang, S., Mai, Z., Xuan, D., and Zhao, W. (2006). Design and Implementation of QoS-Provisioning System for Voice over IP. IEEE Transactions on Parallel and Distributed Systems, vol. 17, no. 3, pp. 276-288.

[13] Myakotnykh, E. S. and Thompson, R. A. (2009). Adaptive Speech Quality Management in Voice-over-IP Communications. Fifth Advanced International Conference on Telecommunications, aict, pp.64-71, Venice/Mestre, Italy.

[14] Raja, A., Azad, R. M. A., and Flanagan, C. (2008). VoIP Speech Quality Estimation in a Mixed Context with Genetic Programming. 10th Annual Conference on Genetic and evolutionary computation, Atlanta, Georgia, United States.

[15] Zha, W. and Chan, W. (2005). Objective Speech Quality Measurement Using Statistical Data Mining. EURASIP Journal on Applied Signal Processing, no. 9, 1410-1424.

[16] Gardner, M., Frost, V.S. and Petr, D.W. (2003). Using optimization to achieve efficient quality of service in Voice over IP networks. IEEE International Performance, Computing, and Communications Conference, Phoenix, Arizona, United States.

[17] Mazurczyk, W. and Kotulski, Z. (2007). Adaptive VoIP with Audio Watermarking for Improved Call Quality and Security. Journal of Information Assurance and Security 2, 226-234. Retrieved from http://www.mirlabs.org/jias/mazurczyk.pdf. Last accessed: 5/6/2010.

[18] Wu, C., Chen, K., Huang, C., and Lei, C. (2009). An Empirical Evaluation of VoIP Playout Buffer Dimensioning in Skype, Google Talk and MSN Messenger. Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital and Video, Williamsburg, VA, United States.

[19] Narbutt, M. and Davis, M. (2005). Assessing the Quality of VoIP Transmission Affected by Playout Buffer Scheme. 4th International Conference on Measurement of Speech and Audio Quality, Prague, Czech Republic.

 ISBN:978-1-61208-113-7

SIP Providers' Awareness of Media Connectivity

Stefan Gasterstädt, Markus Gusowski, Bettina Schnor
Institute of Computer Science
University of Potsdam
Potsdam, Germany
{gasterstaedt,gusowski,schnor}@cs.uni-potsdam.de

***Abstract*—Voice-over-IP (VoIP) has become an important service in the Internet. In contrast to the Public Switched Telephone Network where the delivery of all messages and streams is the responsibility of the calling parties' providers, VoIP media data is sent directly between the user agents without provider interaction in most cases. Hence, a VoIP provider is not aware of media connectivity, i. e., whether a call was successful or not. This may lead to incorrect behavior when a VoIP provider offers services beyond signaling (for example, SPIT prevention, payment). In this paper, we discuss several approaches for the detection of media connectivity and present a solution that conforms with the existing standards. The modified behavior of the user agents, the use of SCTP and provider's awareness of media connectivity are described in detail. Finally, measurements show that our solution results in neglectable overhead.**

***Keywords*-Voice-over-IP (VoIP), media connectivity, SCTP**

I. Introduction

The Session Initiation Protocol (SIP) [22] has become a majorly used protocol in Voice-over-IP (VoIP) communication. It utilizes the Uniform Resource Identifier (URI) schema to address users, single devices or end points and resolves these URIs to Internet Protocol (IP) addresses by using SIP proxy servers and Domain Name Service (DNS) lookups. Users can call others without knowing their current IP address, because session invitations are routed to the SIP proxy that is responsible for the callee's URI domain; and as a next step, this proxy uses its location service to locate the callee[1] and forwards the *INVITE* request to the addressed user (cf. Fig. 1). Depending on its configuration, a SIP proxy may or may not request to stay in the route of any further SIP signaling. Normally, the media transmission is done directly between the user agents (UAs) via RTP.

It is a known problem that the basic SIP infrastructure does not conform to the Network Address Translator (NAT)-friendly application design guidelines described in RFC 3235 [23], and thus, NATs and firewalls cause serious problems for SIP message delivery and media connectivity in conjunction with the separation of signaling and media delivery, dynamic port allocation, or RTP's "$x + 1$" port schema. In contrast to the UA-to-UA media connection, there are solutions for SIP messages; for example, by simply traversing NAT using symmetric response routing [21]. Examples of NAT and firewall traversal for SIP are given in [19].

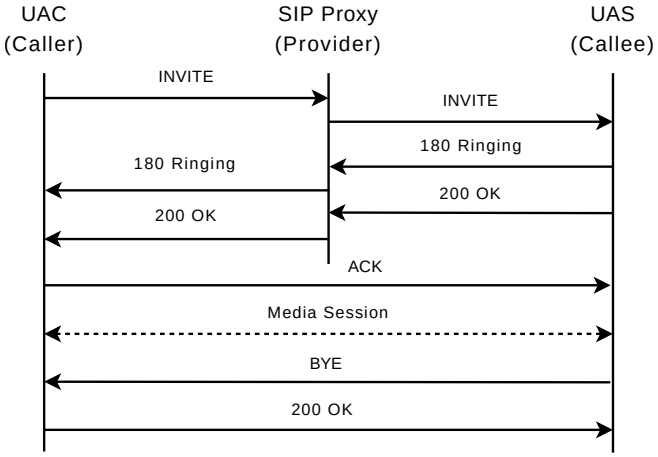


Figure 1: SIP Dialog of a Call

The explicit separation between the session signaling and media delivery comes along with a significant implication: VoIP providers offering SIP services are unaware of whether or not the media stream is actually received by the endpoint(s), i. e., whether there is *connectivity* or not. SIP does not check for connectivity, and the condition is not signaled in any way. Therefore, a SIP provider cannot know if two users will actually be able to communicate, even if a SIP session was successfully established. There are several reasons why media streams negotiated between the UAs may be blocked in one or both directions, mainly because of NATs and/or firewalls [10], [24], but other network problems like the lack of a network route, node crash, configuration problems, or codec mismatch could be responsible as well [1]. This is in contrast to the traditional Public Switched Telephone Network (PSTN), where there is always connectivity once signaling completes successfully.[2]

There are, however, important scenarios where it is desir-

[1]The location bindings can be updated by each respective user sending a *REGISTER* request to its SIP provider's registrar.

[2]Admittedly, there are some rare cases where people cannot talk to each other allthough there has been a successful ringing and call acceptance before. However, the PSTN phone provider will be aware of this failure.

 ISBN:978-1-61208-113-7

able for the provider to know the media connectivity status between the endpoints.

Payment: In some cases, the callee or the caller request some fee in order to accept or initiate a call. Examples include duration-based fees (similar to the PSTN); (fixed) fees relating to the (voice based) service a callee is offering, such as a support hotline; fees for calls a callee subscribed for, such as severe thunderstorm warning; or, in the case of Spam over Internet Telephony (SPIT) prevention, where a caller may be confronted with a small fee if its sincerity is in doubt [11], [12].

For whatever reason a session involves payment by at least one party, it is desirable to delay finalizing the payment transaction until connectivity is assured.

Reputation: Some approaches to detect and prevent SPIT use a reputation score in order to help determine the caller's nature [3], [12], [17]. Each user's reputation is related to its behavior and is calculated from several metrics that are collected by the providers. For examples, a short call duration may indicate an unsolicited call that prompted that callee to hang up immediately. Unfortunately, it may also indicate that at least one participant could not hear the other due to a lack of (bidirectional) media connectivity. In this case, the caller's reputation would falsely be reduced.

Forensics: In the area of law enforcement, reliable evidence is crucial. Regarding the question of whether or not a call took place, SIP can only provide information about signaling – if the phone rang, if the phone was picked up, and if the phone was hung up. This may not be sufficient: The information may be required as to whether or not the two parties in a call were actually able to communicate.

Call Detail Record Analysis: Call Detail Records (CDRs) are collected and analyzed for several reasons. These records contain information about each call, for example, the caller's and callee's IDs, the invitation time, the duration, and how the call terminated. This data can be used to conduct statistical analysis, to profile users' behavior, to reduce traffic congestion or, in general, to detect any kind of anomaly. It is not sufficient if the CDRs are based on the SIP messages only, without knowing whether or not there was media connectivity. This might result in contra-productive network configuration, misinterpretation of someone's reputation or, even worse, will black-list a participant.

In this paper we present a solution for the *VoIP Media Connectivity Awareness Problem*, which fulfills the following requirements:

1) Focus: It is the *SIP Provider* who needs to obtain knowledge about the connectivity status.

2) Multiple (bi-directional) streams: It is important to consider *all media streams* negotiated between the calling parties. Any single uni-directional stream that is not established successfully might be the reason for one of the participant to end the call (immediately). Thus, the provider needs to determine at least the connectivity status for the stream aggregate, with respect to each stream in any direction.

3) Genuineness: In order to prevent false conclusions (and subsequent actions), the connectivity status gathered by the provider should be *genuine*.

4) Compatibility: The number of changes put into the SIP message sequences should be as small as possible. Ideally, neither extra SIP messages nor additional SIP headers should be required.

This paper is structured as follows: In Section II, we discuss several approaches that have some relation to the awareness of media connectivity. In Section III, our approach is presented. This includes detailed scenarios and preliminary investigation of the Stream Control Transmission Protocol (SCTP). Finally, Section IV contains the measurements of the overhead of our solution.

II. Related Work

There are some approaches that relate to the awareness of media connectivity but which are motivated by different goals.

A. Dealing with the NAT

One possibility to solve the connectivity problem is the use of an Application Layer Gateway (ALG) in addition to the NAT. In reality, however, ALGs are deployed in the fewest scenarios, even though most users manage their own private home networks. Furthermore, an ALG might increase the chance to achieve media connectivity, but the SIP provider still does not know about it.

Traversing the NAT for the media streams can be done using Interactive Connectivity Establishment (ICE) [18]. ICE describes NAT traversal for multimedia signaling protocols like SIP, and it extends the Session Description Protocol (SDP) [9] to convey additional data. In order to operate, ICE utilizes the protocols Session Traversal Utilities for NAT (STUN) [20] and Traversal Using Relays around NAT (TURN) [14].

The goal of ICE is to *establish* connectivity, but not to require it or to inform a third party of the connectivity status.

B. Connectivity Preconditions

UAs may use Connectivity Preconditions as defined in RFC 5898 [2] to *verify* whether there is connectivity or not. Based on the concept of a SDP precondition in SIP as specified by RFC 3312 [5] (generalized by RFC 4032 [4]), the connectivity precondition defined by RFC 5898 tries to ensure that session progress is delayed[3] until media stream connectivity has been verified.

Similar to a part of the solution described in this paper (cf. Sec. III), it enables the UAs to delay the SIP session establishment until connectivity is ensured. In contrast to our approach, the provider cannot enforce the UAs to make use

[3]including suppression of alerting the called party

 ISBN:978-1-61208-113-7

of this extension. In addition, it does not inform a third party (such as the provider) of the connectivity status – neither implicitly nor explicitly.

Furthermore, RFC 5898 does not assure that session establishment comes along with media connectivity. In RFC 3312 (which RFC 5898 relates to), alerting the user until all the mandatory preconditions are met has a "SHOULD NOT" semantics.

C. Disconnection Tolerance

Ott and Xiaojun [16] present mechanisms for detection and recovery from temporary service failures for mobile SIP users. For detection of connectivity loss, they suggest a media-based approach: Missing Real-time Transport Protocol (RTP) packets, RTP Control Protocol (RTCP) packets, or STUN packets along with some additional criteria are used as indicators that connectivity has been lost. If the connectivity loss persists longer ("call interruptions"), the UAs will automatically try to re-establish the session after locally terminating the session. For this purpose, the authors introduce the new SIP *Recovery* header field, which is set to `true` in the *INVITE* message used to re-establish the session. The focus of this paper is on obtaining the connectivity status during an ongoing session *after* the session has been established. Implicitly, it assumes that connectivity was given at the beginning of the session.

D. Conclusion

In all solutions presented the focus is always on the endpoints. Whether the main goal is to establish connectivity, ensure connectivity, detect/monitor connectivity status, or recover from connectivity loss, the assumption is always that the *endpoints* are the entities which are interested in the goal.

Hence, the provider is not aware of the media connectivity; and even when the connectivity information can be obtained, its validity and/or genuineness may be questionable.

III. Implicit Connectivity Detection and Notification

One major difference between the approaches presented above is *when* information pertaining to connectivity status is obtained. Three distinct cases can be identified: before session establishment (ICE, Connectivity Preconditions), after session establishment (Disconnection Tolerance [detection only]), and at the end of the conversation (Disconnection Tolerance [signaled through *Recovery* header field]). In the second case, the information can also be obtained continually during the ongoing session.

Another difference is found in the direction of a media stream for which connectivity status is determined and whether media streams are considered separately or jointly on a "session level." Most mechanisms distinguish between individual streams and, as streams are usually considered uni-directional, also between receiving and sending direction. Connectivity Preconditions distinguish both direction and individual streams, but the consequence (suspension of session establishment) is affected by the aggregate of the streams for which the precondition was requested. The Disconnection Tolerance solution disregards direction as symmetric connectivity is assumed; it also disregards individual streams because the existence of only one audio stream is assumed (point-to-point audio conversation).

In our approach, connectivity detection and notification is done before session establishment. Further, our solution regards both, different streams and direction.

A. Implicit Connectivity Notification

SIP itself already offers several possibilities to modify the message routing. For example, a SIP proxy can request to stay in the route of any further SIP messages. Any UA sending a new SIP request needs to insert corresponding routing information. Thus, in contrast to the normal SIP call (see Fig. 1) a proxy can become a mandatory node of the last SIP 3-way-handshake's message (i. e., the *ACK* request). Furthermore, the user agent server (UAS) does not necessarily need to send a *180 Ringing* response and notify the called person. Instead, it can respond with a *183 Session Progress* message to indicate further action prior to call acceptance.

This response message plus the modified message routing can be combined with a modified UA behavior. By using the *183* response's payload, the callee can answer the caller's SDP offer. Thus, both parties know the parameters of all media sessions that normally will be established *after* the SIP session has been accepted. In our solution, the media sessions are established *beforehand*, and both parties **must hold back** the *180 Ringing*, *200 OK*, and the *ACK* messages until this has happened. Furthermore, each UA **must ignore** any incoming media packets as long as the calling partner did not acknowledge the connectivity.

In result, the provider can conclude the media's connectivity status by just analyzing the messages it is routing. Therefore, we call the approach *implicit*. The provider will **conclude that there is connectivity if and only if** the UAS has sent a *200 OK* and then the user agent client (UAC) has sent an *ACK*.

In case the media connection could be established successfully, there will be a notification (*180 Ringing*), acceptance (*200 OK*) and acknowledgement (*ACK*) (see Fig. 2). In result, the provider concludes that there is media connectivity.

If the UAS notices that establishing the media connection failed, it will reject the call by sending a *418* error response (see Fig. 3). If the failure is detected by the UAC (similar to Fig. 4, not shown separately), it will cancel the call using the *CANCEL* request causing the UAS to respond to the

 ISBN:978-1-61208-113-7

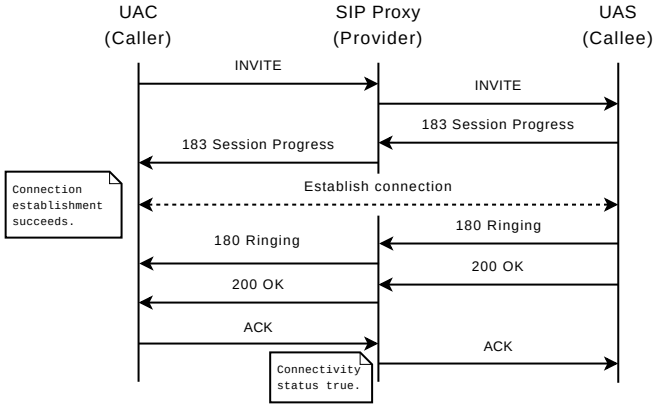


Figure 2: Accepted Call with Prechecked Media Connectivity

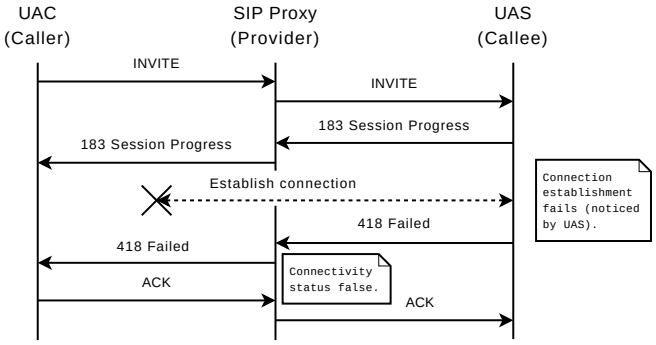


Figure 3: Call Abortion in the Case of no Media Connectivity

invitation with *487 Request Terminated*. In both cases, the provider concludes that there is no media connectivity.

In Figure 4, the media connection has been established successfully but the callee is unavailable. Thus, the caller will cancel the call when a timeout appeared. Again, the provider concludes lack of media connectivity.

B. Detection Connectivity

Due to the fact that the provider is simply analyzing the messages it is routing, it is up to the clients to verify the connectivity status. In detail, they need to check every single media stream for connectivity (cf. Requirement 2). This can be complex and time consuming.

In order to limit this overhead, we propose the use of the Stream Control Transmission Protocol (SCTP) [25] as the media's underlying transport layer. First of all, SCTP is connection oriented; thus, the SCTP's 4-way-handshake at the beginning already ensures transport layer connectivity. In result, neither a media packet nor a notice of receipt need to be sent in order to check for connectivity. Secondly, SCTP itself offers multiplexing; so there is no need for more than one connection, as every single RTP/RTCP stream can be sent using the same unique connection. In result, the time

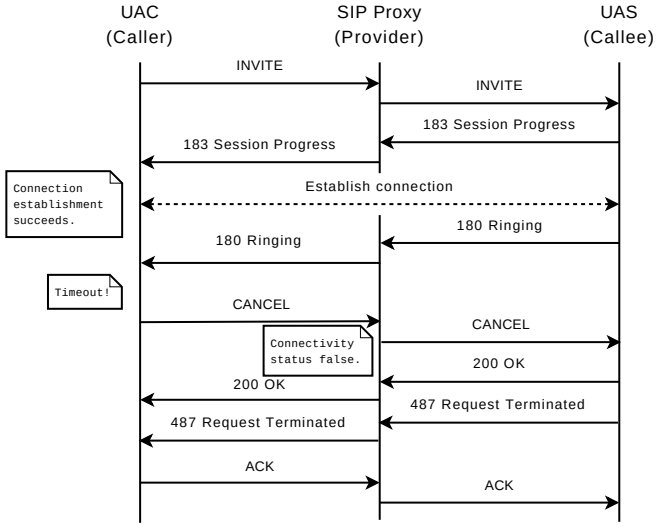


Figure 4: Timed-out Call with Prechecked Media Connectivity

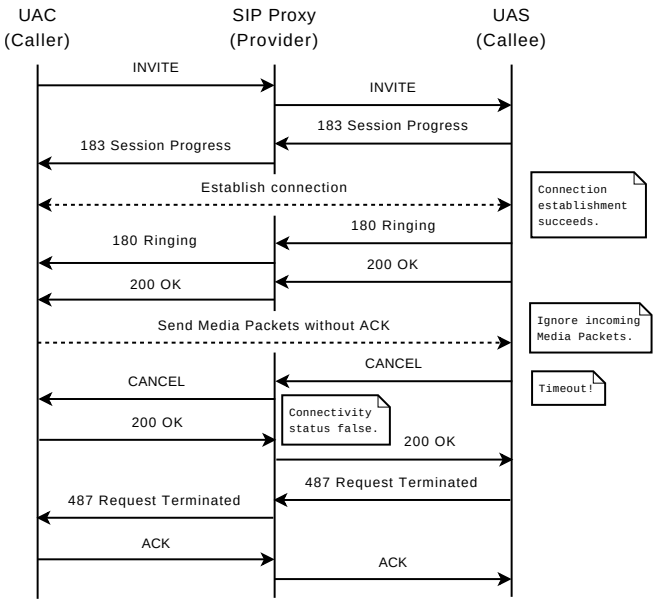


Figure 5: Missbehaving UAC

required to check each media stream (and media control stream) is reduced to a single check only. Last but not least, in contrast to TCP, SCTP offers unordered transport, meaning a lost packet does not delay delivery of succeeding packets. In addition, the partial reliable mode can be used to improve the media quality in case a lost packet can be resent immediately.

To confirm our proposal, we measured the SCTP performance in comparison to UDP. The environment consists of two machines with identical hardware and software running Debian GNU/Linux 5.0.3 (lenny) with kernel version 2.6.26

 ISBN:978-1-61208-113-7

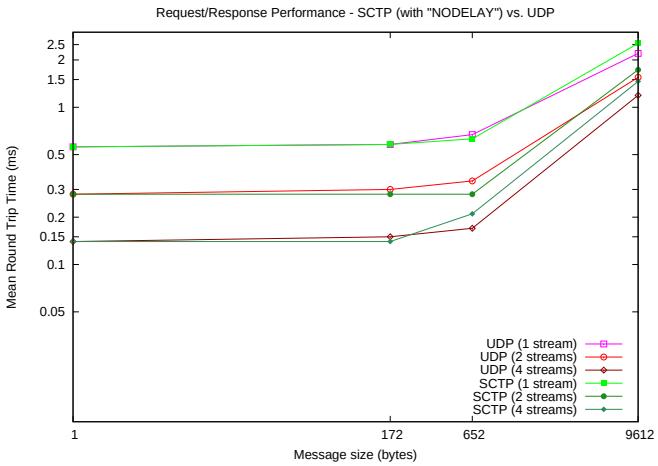


Figure 6: SCTP vs. UDP

(i686). Both machines are equipped with an Intel Core 2 Duo E7500 dual core CPU running at 2.93 GHz and an Intel 82567LM-3 network adapter and connected via a FastEthernet switch (100 Mbps Full Duplex). The environment also determines the UDP and SCTP implementations used – those of the Linux kernel. The benchmark itself is a simple ping-pong application that can send multiple messages at once, approximating multiple concurrent media streams. Figure 6 shows the mean round-trip time (RTT) in relation to the size of the messages. The sizes of 172 Bytes and 652 Bytes correlate to the RTP packet sizes produced by the G.711 codec using packet transmission cycles of 20 ms and 80 ms, respectively. One can see that the values of SCTP are quite similar to UDP, and hence, we expect no performance loss due to the use of SCTP.

C. Missbehaving user agents

In some cases, either the UAS or the UAC might try to falsify the information it tells about the connectivity status. Our solution would require to send a *200 OK* (UAS) or *ACK* (UAC) to convey "connectivity."

For example, a caller might announce "no connectivity" to the provider in order to send SPIT calls without consequences. In our solution, the UAC would have to suppress the *ACK* message. Fortunately, this would cause the callee to ignore any incoming media packet (cf. Fig. 5). In the payment example, the callee might say "connectivity" in order to receive his fee anyway. In this case, the caller receives a *200 OK* even though there is no media connectivity. In result, he can abort the call by sending a *CANCEL* request. In general, for whatever reason a UA might missbehave – our solution enables the other party to react appropriately, enabling the provider to know the actual connectivity status.

The case that both, caller and callee, are lying cooperatively, this is only a problem in the forensics scenario. It is doubtful, however, that the calling partners would use a provider at all in such a scenario.

D. Protocol Extensions

There has been some work in the past for SCTP and SIP, but unfortunately, it is incomplete and has been abandoned [6], is limited in its scope [13], and does not deal with the use of RTP over SCTP.

In line with the last requirement, our solution only needs to slightly extend the abilities of SDP in order to specify the SCTP parameters. The use of the SCTP connection and the modified UA behavior can be indicated by naming our extension (i. e., *sctp-tunnel*) within a *Require* header. If an incoming *INVITE* does not indicate usage of this extension the provider must reject this request by sending *421 Extension Required*. As described above, the extension just specifies the way the UAs must behave and the provider can draw conclusions; it does not specify any new SIP messages or headers – all of them have existed before. The syntactical details of both modifications can be found in [8].

IV. Measurements

Although we minimized the changes to the existing VoIP infrastructure, the provider still has to be aware of the *sctp-tunnel* extension indicated within the SIP messages. Whether the extension is stated or not, the provider has to use different message handling and routing. It is thus important to know how much overhead our extension creates.

Note that the following measurements do not cover the impact of SCTP. SCTP is used as the underlying protocol of the UA-to-UA media session only, whereas SIP messages still use UDP. In result, a SIP proxy does not need to be adapted to use another transport protocol. On the other hand, media gateways (not considered by the following measurements) need to be altered to conform to our approach.

A. Testbed, Scenarios

We used three nodes (each with 2 x AMD Opteron 244 CPU (1.8 GHz), 4 GB RAM, Gigabit Ethernet Interconnection) to setup one SIP proxy (Kamailio [15], v3.0.3) and two UAs (SIPp [7], v3.1) that generated and processed a various number of SIP calls. Kamailio has been configured to use 1024 MB of memory, to create four processes, and its log level was set to zero.

In general, we measured three scenarios: a) default behavior of the proxy, b) modified behavior of the proxy where the UAs already indicated the use of the *sctp-tunnel* extension, and c) the modified behavior of the proxy without initial indication by the UAs. The third scenario is the most expensive one since the provider needs to reject incoming invitations first, and then has to deal with the reformulated ones. In addition, we measured d) the SIPp-SIPp-interconnectivity to determine the overhead of Kamailio in general.

In scenarios a) and b), the UAC and the UAS send and receive SIP messages according to Figure 7. In contrast to

 ISBN:978-1-61208-113-7

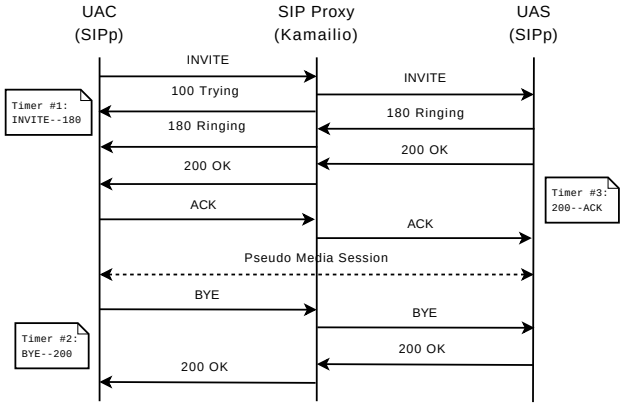


Figure 7: Measurement Scenario

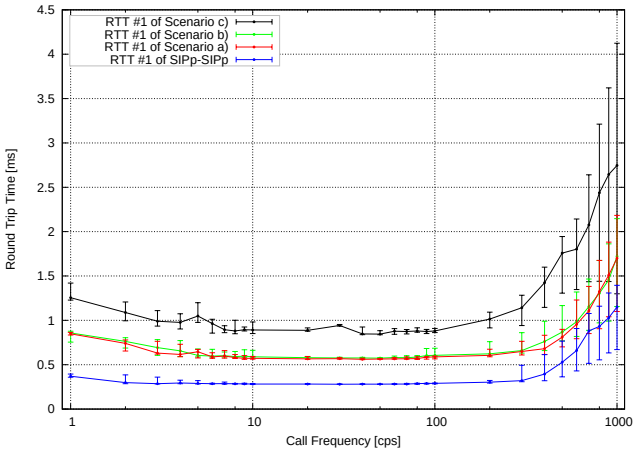


Figure 8: Comparison of RTT #1

the example given in Figure 1, the proxy stays in the route for the whole call. Not shown separately, scenario c) requires three more messages at the beginning: the first *INVITE* will be rejected with a *421* response that has to be *ACK*ed.

Each call generates three round-trip time values: RTT #1 represents the delay of a UAS's response including Kamailio action (such as lookup and extension verification); RTT #2 represents the delay of a UAS's response in case the request can be forwarded immediately; RTT #3 represents the delay of a UAC's feedback. Each series lasted five minutes, using a constant call frequency (between 1 and 1000 calls per second). The proxy and the UAs were restarted for each frequency.

B. Results

For all scenarios and each frequency we calculated the corresponding median and quartile values for each RTT. As expected, in scenarios a)–c), the values of RTT #2 and RTT #3 are nearly the same. The SIPp-SIPp interconnection's second and third RTT are ∼0.25–0.55 ms lower only. RTT #2 and #3 are not shown separately since our proposal does not alter the way *200 OK* and *ACK* messages or session tear down is handled.

The comparison of RTT #1 is shown in Figure 8. Again, one can see the additional time required (∼0.5–0.6 ms) when Kamailio is put between the SIPp instances. Furthermore, we expected the overhead of the header verification to be very small since we only slightly modified the routing logic of Kamailio. This small RTT increase can be seen when comparing the values of scenarios a) and b).

In scenario c), where the proxy had to enforce the use of the SIP extension, RTT #1 increases a little more. This happens because Kamailio is involved one more time and three more messages are sent until the first callee's response is received by the inviting caller.

V. Conclusion

In this paper, we have given several scenarios motivating the need of SIP providers' awareness of media connectivity, such as payment, reputation, forensics, and call detail record analysis.

In our solution, the provider is *implicitly* informed about the media connectivity: The SIP provider can draw genuine conclusions by simply analyzing the messages it is routing. The UA, however, needs to alter its behavior. This behavior is specified by way of a new SIP extension and its usage can be enforced by the provider.

To reduce the overhead of media connectivity detection, we propose to use SCTP for media transport. This requires a slight extension of SDP.

The measurements have shown that the overhead induced by our solution is neglectable, as long as the UAs indicate the use of our extension from the beginning. In addition, our approach can easily be integrated into existing VoIP infrastructures as it fully conforms to existing protocols. If a UA is not aware of our extension it is at the discretion of the provider to proceed with the call (without the ability to conclude media connectivity) or to reject it.

Future work will deal with Quality of Service (QoS) aspects. Besides a lack of connectivity, low quality can also cause a call to be aborted prematurely by one of the participants. We therefore need to conduct further investigation in order to deal with this problem.

References

[1] Alessandro Amirante, Simon Pietro Romano, Kyung Hwa Kim, and Henning Schulzrinne. Online Non-Intrusive Diagnosis of One-Way RTP Faults in VoIP Networks Using Cooperation. In Georg Carle, Helmut Reiser, Gonzallo Camarillo, and Vijay K. Gurbani, editors, *Proceedings of the 4th International Conference on Principles, Systems and Applications of IP Telecommunications (IPTComm 2010)*,

pages 153–160, Munich, Germany, August 2^{nd}–3^{rd}, 2010. Technical University Munich.

[2] F. Andreasen, G. Camarillo, D. Oran, and D. Wing. Connectivity Preconditions for Session Description Protocol (SDP) Media Streams. RFC 5898 (Proposed Standard), July 2010.

[3] Vijay A. Balasubramaniyan, Mustaque Ahamad, and Haesun Park. CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation. In *Proceedings of the 4^{th} Conference on Email and AntiSpam, CEAS 2007*, August 2^{th}–3^{rd}, 2007.

[4] G. Camarillo and P. Kyzivat. Update to the Session Initiation Protocol (SIP) Preconditions Framework. RFC 4032 (Proposed Standard), March 2005.

[5] G. Camarillo, W. Marshall, and J. Rosenberg. Integration of Resource Management and Session Initiation Protocol (SIP). RFC 3312 (Proposed Standard), October 2002. Updated by RFCs 4032, 5027.

[6] R. Fairlie-Cuninghame. Guidelines for specifying SCTP-based media transport using SDP. Internet-Draft draft-fairlie-mmusic-sdp-sctp-00, Internet Engineering Task Force, May 2001. Work in progress.

[7] Richard Gayraud, Olivier Jacques, et al. SIPp: An Open Source Performance Testing Tool for SIP [v3.1]. http://sipp.sourceforge.net, March 17^{th}, 2009.

[8] Markus Gusowski. Media Connectivity in VoIP Infrastructures: Requirements, Detection, Enforcement. Master's thesis, University of Potsdam, July 2010.

[9] M. Handley, V. Jacobson, and C. Perkins. SDP: Session Description Protocol. RFC 4566 (Proposed Standard), July 2006.

[10] M. Holdrege and P. Srisuresh. Protocol Complications with the IP Network Address Translator. RFC 3027 (Informational), January 2001.

[11] C. Jennings, J. Fischl, H. Tschofenig, and G. Jun. Payment for Services in Session Initiation Protocol (SIP). Internet-Draft draft-jennings-sipping-pay-06, Internet Engineering Task Force, July 2007. Work in progress.

[12] S. Liske, K. Rebensburg, and B. Schnor. SPIT-Erkennung, -Bekanntgabe und -Abwehr in SIP-Netzwerken. In U. Ultes-Nitsche, editor, *Proceedings of KiVS – NetSec 2007, Workshop „Secure Network Configuration“*, pages 33–38, February 2007.

[13] S. Loreto and G. Camarillo. Stream Control Transmission Protocol (SCTP)-Based Media Transport in the Session Description Protocol (SDP). Internet-Draft draft-loreto-mmusic-sctp-sdp-05, Internet Engineering Task Force, February 2010. Work in progress.

[14] R. Mahy, P. Matthews, and J. Rosenberg. Traversal Using Relays around NAT (TURN): Relay Extensions to Session Traversal Utilities for NAT (STUN). RFC 5766 (Proposed Standard), April 2010.

[15] Ramona-Elena Modroiu, Bogdan Andrei Iancu, Daniel-Constantin Mierla, et al. Kamailio (OpenSER) [v3.0.3]. http://www.kamailio.org/, August 19^{th}, 2010.

[16] Jörg Ott and Lu Xiaojun. Disconnection tolerance for SIP-based real-time media sessions. In *MUM '07: Proceedings of the 6^{th} international conference on mobile and ubiquitous multimedia*, pages 14–23, New York, NY, USA, 2007. ACM.

[17] J. Rosenberg. The Session Initiation Protocol (SIP) and Spam. Internet-Draft draft-ietf-sipping-spam-03, Internet Engineering Task Force, October 2006. Work in progress.

[18] J. Rosenberg. Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols. RFC 5245 (Proposed Standard), April 2010.

[19] J. Rosenberg and G. Camarillo. Examples of Network Address Translation (NAT) and Firewall Traversal for the Session Initiation Protocol (SIP). Internet-Draft draft-rosenberg-sipping-nat-scenarios-03, Internet Engineering Task Force, July 2004. Work in progress.

[20] J. Rosenberg, R. Mahy, P. Matthews, and D. Wing. Session Traversal Utilities for NAT (STUN). RFC 5389 (Proposed Standard), October 2008.

[21] J. Rosenberg and H. Schulzrinne. An Extension to the Session Initiation Protocol (SIP) for Symmetric Response Routing. RFC 3581 (Proposed Standard), August 2003.

[22] 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), June 2002. Updated by RFCs 3265, 3853, 4320, 4916, 5393, 5621, 5626, 5630, 5922.

[23] D. Senie. Network Address Translator (NAT)-Friendly Application Design Guidelines. RFC 3235 (Informational), January 2002.

[24] P. Srisuresh and K. Egevang. Traditional IP Network Address Translator (Traditional NAT). RFC 3022 (Informational), January 2001.

[25] R. Stewart. Stream Control Transmission Protocol. RFC 4960 (Proposed Standard), September 2007.

 ISBN:978-1-61208-113-7

Improving SIP authentication

Lars Strand

Norwegian Computing Center / University of Oslo
Oslo, Norway
Email: lars.strand@nr.no

Wolfgang Leister

Norwegian Computing Center
Oslo, Norway
Email: wolfgang.leister@nr.no

***Abstract*—The digest access authentication method used in the voice over IP signaling protocol, SIP, is weak. This authentication method is the only method with mandatory support and widespread adoption in the industry. At the same time, this authentication method is vulnerable to a serious real-world attack. This poses a threat to VoIP industry installations and solutions. In this paper, we propose a solution that counters attacks on this wide-spread authentication method.**

***Index Terms*—SIP, authentication, Digest Access Authentication, security attack.**

I. Introduction

The most common protocol pair used for sending Voice over IP (VoIP) is the Session Initiation Protocol (SIP) [1] and Real-time Transport Protocol (RTP) [2]. RTP transfers the media content, while SIP handles the signaling, i.e., set up, modification and termination of sessions between two or more participants. VoIP is the emerging technology that will eventually take over from the traditional Public Switched Telephone Network (PSTN) [3] due to VoIP's improved flexibility and functionality, such as improved sound quality ("HD sound") using wideband codecs like G.722 [4], instant messaging (IM), presence, mobility support, and secure calls. VoIP reduces maintenance and administration costs since it brings convergence to voice, video and data traffic over the IP infrastructure.

SIP is an application layer protocol developed by the IETF. Its core functionality is specified in RFC3261 [1]. Additional functionality is specified in additional RFCs [5]. SIP sessions range from ordinary calls between two participants to advanced conference sessions between multiple participants communicating over video, voice, and IM.

However, SIP and RTP-based VoIP installations are rather difficult to secure [6]. VoIP inherits many security threats and Quality of Service (QoS) properties from the Internet, in addition to threats that come from the VoIP-specific technologies [7]. A clear and concise VoIP threat taxonomy is given by VOIPSA [8]. There are many obstacles in securing SIP, due to its use of intermediaries and the fact that functionality was the primary focus for the SIP designers, not security [1, page 232].

SIP supports several security services, and the RFC recommends their use. These security services can provide protection for authentication, confidentiality, and more. Yet, only one such security service is mandatory: the SIP Digest Access Authentication (DAA) method [1, page 193]. In our experience the other security services are neither implemented nor used. The only security service used is the mandatory authentication method.

DAA is primarily based on the HTTP Digest Access Authentication [9], and is considered to be weak and vulnerable to serious real-world attacks [10].

The main contribution of this paper is to present and analyze the seriousness of a vulnerability we presented in our earlier work – the registration attack [10]. We propose a solution to secure DAA that will counter this vulnerability.

The rest of the paper is organized as follows: We show our approach in Section II. We explain SIP authentication in Section III, and show the registration attack previously discovered in Section IV. In Section V, we show how to improve the authentication method to counter this attack. Related work is given in Section VI, before concluding in Section VII.

II. Method and case study

In Norway, both private companies and public authorities are migrating from PSTN to VoIP [11]. Our case study is taken from three companies in Norway; one medium sized company with 150 employees, and two larger companies with 3000 and 4700 employees. We have gathered several of these VoIP configurations and setups, and replicated the installations in our test lab [12]. In these companies, most of the employees have their own VoIP phone, called a User Agent (UA). All VoIP servers run the Linux operating system with the open source telephony platform *Asterisk* [13]. We found in these configurations that the digest authentication is the only authentication method for the UAs.

Our analysis follows the workflow shown in Fig. 1. In the following paragraphs, the numbers in parentheses refer to the numbers in Fig. 1.

In order to gain knowledge of the SIP protocol we use the specification documents (1), here the SIP standard. Then, we analyze VoIP network traffic going through the test lab (5). We have implemented two VoIP setups based on configurations from our industry partners ((2) and (3)). The network traffic is intercepted and saved to file using the network tool *tcpdump* (4). The network traffic is then analyzed off-line using the packet analyzer, *Wireshark* (5). An example of such an analysis is shown in Fig. 2.

 ISBN:978-1-61208-113-7

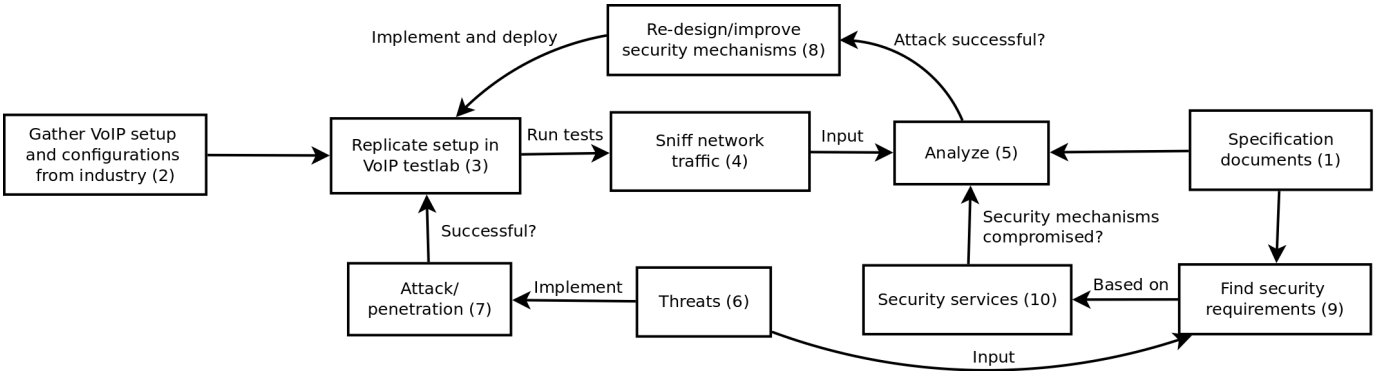


Fig. 1: Workflow for analysis of the SIP authentication method.

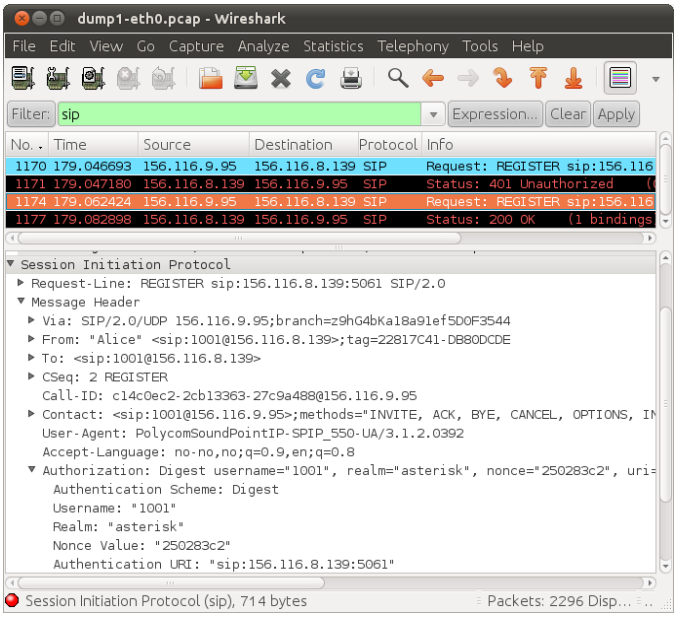


Fig. 2: Network analysis using the network tool Wireshark.

As an additional input we consider threats deducted from formal analysis of the protocol, such as a SIP attack analyzed by Hagalisletto and Strand [10], using the protocol analyzer PROSA (6). We explain the attack in more detail in Section IV, and implement and execute the attack using the network tool *NetSED* (7) as shown in Fig. 6. Based on the security requirements (9) obtained from the SIP specification, we then checked if the authentication method (10) was compromised by the real-world attack. After careful analysis of the SIP headers we found that the SIP registration attack could be countered by a modification of the SIP authentication method (8).

III. Authentication in SIP

Authentication is the assurance that a communicating entity is the one that it claims to be [14]. Authentication consists of two basic steps: *a*) *Identification*, where an entity/client presents a value to the authentication system, and *b*) *Verification* where this value is validated against the authentication system [15]. When people that know each other are dialing or answering a phone call, they can often authenticate the other by just recognizing the other person's voice. However, when using new communications channels, such as instant messaging (IM), video, screencast and presence, determining the authenticity of the communicating partner is more difficult than for a voice call. To have established the identity of the caller is also important when, for instance, a physician need to communicate with a patient and discuss sensitive health information. For instance, someone else could masquerade as the patient and illegally obtain sensitive health information on the patient.

The SIP Digest Access Authentication (DAA) is currently the most common authentication scheme for SIP. Other authentication schemes have emerged, but DAA is the only mandatory authentication scheme [1, Section 22]. DAA uses a challenge-response pattern, and relies on a shared secret between client and server.

SIP is heavily influenced by the HTTP request-response model, where each transaction consists of a request that requires a particular response. The SIP messages are also similar in syntax and semantics to both HTTP and SMTP [16]. A SIP message consists of headers and a body. The SIP header fields are textual, always in the format `<header_name>: <header_value>`. The header value can contain one or more parameters. We show an example SIP header message in Fig. 4.

Any SIP request can be challenged for authentication. We show an example SIP DAA handshake in Fig. 3, and refer to the protocol clauses with a number in parentheses. The initial SIP `REGISTER` message (1) from Alice is not authorized and must be authenticated. The SIP server responds with a `401 Unauthorized` status message (3) which contains a `WWW-Authenticate` header with details of the challenge, including a *nonce* value. The client computes the required SIP digest that is embedded in (4) as an `Authorization` header. The SIP server, upon receiving the `Authorization` header, must perform the same digest operation, and compare the result. If the results are identical, the client is authenticated,

 ISBN:978-1-61208-113-7

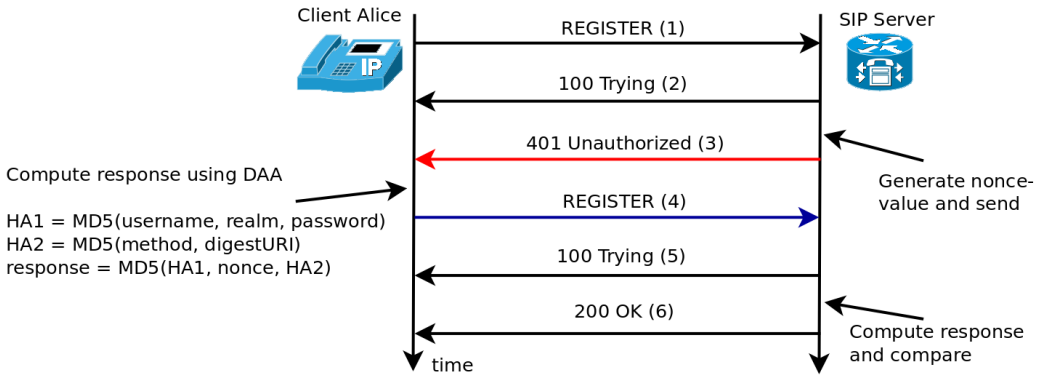


Fig. 3: The SIP Digest Access Authentication method during a SIP `REGISTER` transaction.

```
1.  REGISTER sip:CompanyA SIP/2.0
2.  Via: SIP/2.0/UDP
    156.116.9.95;branch=z9hG4bK32F3EC44EB23347BFB0D488459C69E4E
3.  From: Alice <sip:alice@CompanyA>;tag=1234648905
4.  To: Alice <sip:alice@CompanyA>
5.  Contact: "Alice" <sip:alice@156.116.9.95:5060>
6.  Call-ID: 2B6449C74C10D4F95006A6C034E79E8E@CompanyA
7.  CSeq: 19481 REGISTER
8.  User-Agent: PolycomSoundPointIP-SPIP_550-UA/3.1.2.0392
9.  Authorization: Digest
    username="alice",realm="asterisk",nonce="3b7a1395",response="
    ccbde1c3c129b3dcaa14a4d5e35519d7",uri="sip:CompanyA",algorith
    m=MD5
10. Max-Forwards: 70
11. Expires: 3600
12. Content-Length: 0
```

Fig. 4: The only attributes included in the digest response (blue) are depicted in green.

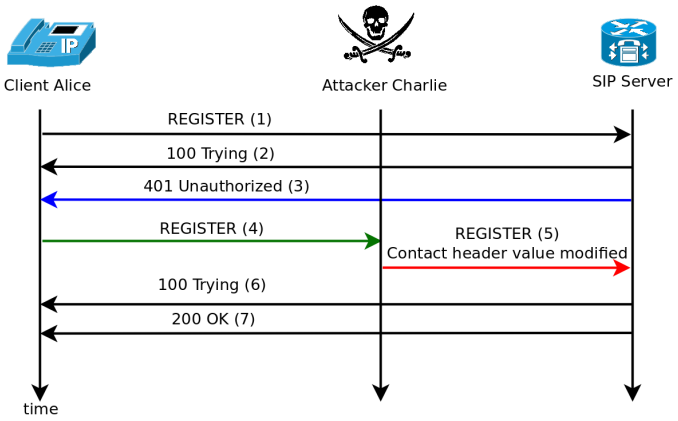


Fig. 5: The attacker Charlie can modify the `Contact` header value, and thereby have all Alice's calls redirected to him.

and a `200 OK` message (6) is sent.

The SIP DAA is almost identical to the HTTP digest access authentication [9]. As we will show later, too few attributes are included in the digest computation, thus leaving some values unprotected. Formally, the DAA is expressed as follows:

$$\begin{aligned} HA1 &= \mathrm{MD5}(A1) \\ &= \mathrm{MD5}(username : realm : password) \\ HA2 &= \mathrm{MD5}(A2) = \mathrm{MD5}(method : digestURI) \\ response &= \mathrm{MD5}(HA1 : nonce : HA2) \end{aligned}$$

In this context, *A1* is the concatenated string of Alice's *username*, the *realm* (usually a hostname or domain name) and the shared secret *password* between Alice and the server. For *A2*, the *method* is the SIP method used in the current transaction, in the above example that would be `REGISTER`. In a `REGISTER` transaction the *digestURI* is set to the URI in the *To:*-field. The digest authentication *response* is the hash of the concatenated values of *HA1*, the *nonce* received from the server, and *HA2*. A SIP `REGISTER` message with a computed digest embedded in the `Authorization` header is shown in Fig. 4. DAA provides only reply protection due to the nonce value and one-way message authentication. There is no encryption of the content, nor confidentiality support, except the shared secret *password* between client and server. All messages are sent in clear. DAA only works within a local domain so cross-domain authentication is not supported, which implies that end-to-end authentication is not supported. There is no provision in the DAA for the initial secure arrangement between a client and server to establish the shared secret. However, DAA has low computation overhead compared to other methods [17].

IV. Attack on DAA

When a UA comes online it registers its contact point(s) to a *location service*. Contact points are the preferred methods a user can be contacted by, for example using SIP, mail, or IM. Usually, only a SIP URI contact method is present. The location service is responsible to redirect SIP requests (for VoIP calls) to the correct SIP end-point. For example, an incoming SIP call destined to `alice@CompanyA.org` does not contain information about which hostname or IP-address Alice's phone can be reached. Therefore, a SIP proxy will query the location service to receive Alice's phone's hostname or IP-address, and then redirect the call to this address.

The binding of Alice's phone to a hostname or IP-address is done during the `REGISTER` transaction, as depicted in Fig. 3. Before the binding, or registration, the SIP server should ask the client to authenticate itself, as explained in the

 ISBN:978-1-61208-113-7

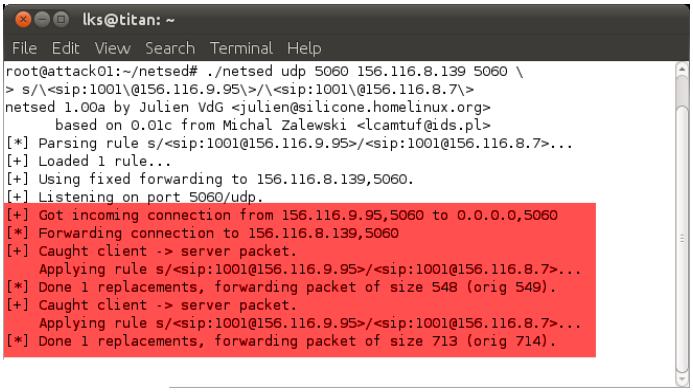


Fig. 6: The network packet stream editor NetSED modifies network packets in real time based on a regular expression (in red).

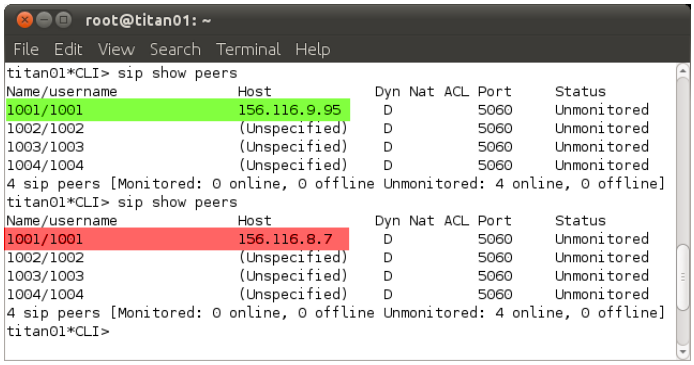


Fig. 7: Host name before (green) and after a successful attack (red), which makes Asterisk believe that Alice's phone (with number 1001) is reachable at an IP-address of the attacker's choice.

previous section. After a successful authentication, the client's hostname or IP-address is registered. A re-registration is normally done at regular intervals. This registration is repeated usually every 3-10 minutes, depending on the configuration. The client's preferred contact methods, including hostname or IP-address, is carried in the SIP header `Contact`, as depicted in Line 5 in Fig. 4. However, this SIP header value is sent in clear, and is not protected by DAA. Thus, the registration is vulnerable to a man-in-the-middle attack [10].

If an attacker modifies the hostname or IP-address in the $contactURI$ header value during a `REGISTER` phrase, as depicted in Fig. 5, all requests, and hence calls, to the client will be diverted to a hostname or IP-address controlled by an attacker. Here, Alice cannot perceive that she is unreachable. An attacker can modify Alice's `REGISTER` session in real-time using NetSED [18] as depicted in Fig. 6. The SIP server (Asterisk), will not detect nor suspect that anything is wrong, and register Alice's phone number with the attackers IP address, as seen on Asterisk's terminal in Fig. 7. When Asterisk receives a call to Alice, the call will be forwarded to the attackers registered IP address.

V. Improving DAA

The SIP digest authentication is weak, which is stated in both the SIP specification [1], and the digest specification [9]. Specifically, DAA only offers protection of the value in the `To` header called the `Request-URI` and the $method$, but no other SIP header values are protected. Other better and stronger authentication methods have been recommended [19]. Nonetheless, we suggest improving the DAA as well as possible, since DAA is the authentication method commonly used due to its simplicity and widespread support and adoption.

A minor modification of DAA can counter the registration hijack attack [10], which is caused by having too few SIP header parameters protected by the digest. Since an attacker can modify and redirect all requests, we protect the header by including the `Contact` header value in the digest. By including the `Contact` value, which we name $contactURIs$ in the digest, we effectively counter the registration hijack attack.

We define $HA0$ with $contactURIs$. The new digest computation algorithm is as follows:

$$\begin{aligned} HA0 &= \text{MD5}(A0) = \text{MD5}(contactURIs) \\ HA1 &= \text{MD5}(A1) \\ &= MD5(username : realm : password) \\ HA2 &= \text{MD5}(A2) = \text{MD5}(method : digestURI) \\ response &= \text{MD5}(HA0 : HA1 : nonce : HA2) \end{aligned}$$

Weaknesses in the MD5 hash have been found. In particular we mention collision attacks where two different input values produce the same MD5 hash [20]. This weakness is not known to be exploitable to reveal a user's password [21]. Nonetheless, a stronger hash function, like SHA1 [22], is recommend.

We implemented and tested our modified DAA by using the Python Twisted [23] networking engine, using both MD5 and SHA1. According to our test, the computation overhead by including $HA0$ with the $ContactURIs$ is minimal, as shown in Fig. 8. The difference between the original DAA and our modified DAA with MD5 for 100.000 authentication requests on a 2.2Ghz Intel CPU, is only 0.44 seconds, a negligible amount.

A modified DAA means a modification of the SIP standard. Since the SIP standard has seen widespread industry adoption, it can be difficult to re-deploy a non-standardized SIP DAA. To prevent a modification of the SIP standard, we can use the DAA parameter `auth-param` to store our modified digest response. The parameter `auth-param` is reserved "for future use" [9, page 12], and can be a part of the `Authorization` header.

SIP devices that do not support the updated and more secure digest, can and will ignore this value, and use the original DAA for authentication. However, we cannot recommend this approach, since an attacker could remove this value and force the usage of the original standardized DAA. We would prefer to modify the DAA digest computation to force an upgrade to

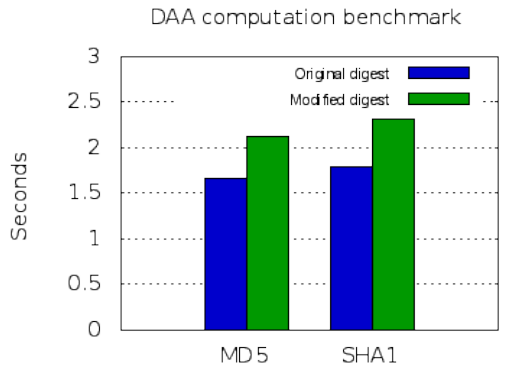


Fig. 8: The computation overhead for 100.000 iterations for original DAA and our modified DAA for both MD5 and SHA1.

the new improved DAA method, instead of compromising on security.

VI. Related work

Based on the DAA, Undrey [24] proposed a more flexible use of variables protected by the digest. His paper addresses the shortcomings of DAA and suggests to allow the server to decide which headers it requires to be included and protected by the digest computation. Unfortunately, his approach does not require specific headers fields to be included. Therefore, transactions that do not include `Contact` fields are still vulnerable to the registration attack.

Palmieri et al. [25], [26], dismiss DAA as a usable authentication method, and instead craft a new authentication schema with digital signatures based on public-key encryption. They rely on public key infrastructure (PKI), but admit that PKI is difficult and costly to implement.

Yang et al. [27] also conclude that DAA is weak. They argue that, since DAA is vulnerable to an off-line password guessing attacks, a more secure authentication method is required. They propose an authentication method based on Diffie-Hellman. Unfortunate, they do not discuss nor add any additional SIP headers in their new authentication scheme. So their solution is also vulnerable to the registration attack.

The H.323 recommendation for the VoIP protocol from the International Telecommunication Union (ITU) has failed to see widespread adoption by industry players, and is considered abandoned in favor of SIP/RTP [16]. The authentication methods in H.323, specified in H.235 [28], [29] uses well established security mechanism, like certificates, and Diffie-Hellman key exchange, to enforce authentication. Further analysis is needed to see whether the H.235 standard protects the signaling better than SIP.

The Inter-Asterisk eXchange (IAX) [30], also published by the IETF, establishes a competing protocol to SIP/RTP. IAX has several security properties that are better than SIP. By multiplexing channels over the same link and transporting both signaling and media over the same port, enforcing security mechanisms is easier. IAX supports two authentication methods: *1*) MD5 Message Digest authentication [31] computed over a pre-shared secret and a challenge (nonce), or *2*) using RSA public-key encryption on the challenge. In both methods, the nonce value is the only protocol parameter that is integrity protected by the authentication. Future work needs to investigate whether the IAX authentication method is adequately secure.

Other, more secure, authentication methods for SIP have been standardized, such as the support for public key encryption with S/MIME [32], the "Asserted Identity" extension [33], and the "Identity" header extension [34]. None of these authentication methods have seen any widespread deployment yet [19].

VII. Conclusion

We have seen that the widely deployed authentication method DAA in SIP is weak and vulnerable to attacks. Moreover, we have confirmed and verified that the attack analyzed earlier [10] can be performed on the SIP protocol in real-time. We have examined this authentication method, and proposed a solution to counter the serious registration attack. By including more SIP header parameters in the authentication digest this attack can be countered.

The original SIP designers focused on functionality and compliance at the cost of security. A more thorough investigation of the SIP DAA in the design phase would have revealed the vulnerability presented here, and the vulnerability could have been prevented early on.

Our remedy presented here solves an serious problem with the DAA. However, other weaknesses and shortcomings of DAA are too serious to be part of a strong and secure authentication scheme for SIP. Therefore, we intend to investigate other authentication methods for SIP, including support for Generic Security Service API (GSS-API) [35].

Acknowledgment

This research is funded by the EUX2010SEC project in the VERDIKT framework of the Norwegian Research Council (Norges Forskningsråd, project 180054). The authors would like to thank Trenton Schulz for discussions. We also thank Anders Moen Hagalisletto and the anonymous reviewers for comments on earlier drafts of this paper.

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, updated by RFCs 3265, 3853, 4320, 4916, 5393, 5621, 5626, 5630, 5922, 5954, 6026. [Online]. Available: http://www.ietf.org/rfc/rfc3261.txt [Accessed: 1. Nov 2011]

[2] 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, updated by RFCs 5506, 5761. [Online]. Available: http://www.ietf.org/rfc/rfc3550.txt [Accessed: 1. Nov 2011]

[3] L. Strand and W. Leister, "A Survey of SIP Peering," in NATO ASI - Architects of secure Networks (ASIGE10), May 2010.

[4] International Telecommunication Union, "7 kHz Audio-Coding within 64 kbits/s," ITU-T Recommendation G.722, 1993.

 ISBN:978-1-61208-113-7

[5] "IETF Session Initiation Protocol Core Charter." [Online]. Available: http://datatracker.ietf.org/wg/sipcore/charter/ [Accessed: 1. Nov 2011]

[6] D. York, Seven Deadliest Unified Communications Attacks. Syngress, Apr. 2010.

[7] H. Dwivedi, Hacking VoIP: Protocols, Attacks, and Countermeasures, 1st ed. No Starch Press, Mar. 2009.

[8] VoIPSA, "VoIP security and privacy threat taxonomy," Public Realease 1.0, Oct. 2005. [Online]. Available: http://voipsa.org/Activities/VOIPSA_Threat_Taxonomy_0.1.pdf [Accessed: 1. Nov 2011]

[9] J. Franks, P. Hallam-Baker, J. Hostetler, S. Lawrence, P. Leach, A. Luotonen, and L. Stewart, "HTTP Authentication: Basic and Digest Access Authentication," RFC 2617 (Draft Standard), Internet Engineering Task Force, Jun. 1999. [Online]. Available: http://www.ietf.org/rfc/rfc2617.txt [Accessed: 1. Nov 2011]

[10] A. M. Hagalisletto and L. Strand, "Formal modeling of authentication in SIP registration," in Second International Conference on Emerging Security Information, Systems and Technologies SECURWARE '08. IEEE Computer Society, August 2008, pp. 16–21.

[11] L. Fritsch, A.-K. Groven, L. Strand, W. Leister, and A. M. Hagalisletto, "A Holistic Approach to Open Source VoIP Security: Results from the EUX2010SEC Project," International Journal on Advances in Security, no. 2&3, pp. 129–141, 2009.

[12] L. Strand, "VoIP lab as a research tool in the EUX2010SEC project," Norwegian Computing Center, Department of Applied Research in Information Technology, Tech. Rep. DART/08/10, April 2010.

[13] "Asterisk: The Open Source PBX & Telephony Platform." [Online]. Available: http://www.asterisk.org/ [Accessed: 1. Nov 2011]

[14] International Telecommunication Union (ITU), "Security Architecture For Open Systems Interconnection (OSI)," The International Telegraph and Telephone Consultative Comittee (CCITT), X.800 Standard, 1991.

[15] R. Shirey, "Internet Security Glossary, Version 2," RFC 4949 (Informational), Internet Engineering Task Force, Aug. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4949.txt [Accessed: 1. Nov 2011]

[16] H. Sinnreich and A. B. Johnston, Internet communications using SIP: Delivering VoIP and multimedia services with Session Initiation Protocol, 2nd ed. New York, NY, USA: John Wiley & Sons, Inc., August 2006.

[17] S. Salsano, L. Veltri, and D. Papalilo, "SIP security issues: The SIP authentication procedure and its processing load," Network, IEEE, vol. 16, pp. 38–44, 2002.

[18] "NetSED: The network packet stream editor." [Online]. Available: http://silicone.homelinux.org/projects/netsed/ [Accessed: 1. Nov 2011]

[19] D. Sisalem, J. Floroiu, J. Kuthan, U. Abend, and H. Schulzrinne, SIP Security. WileyBlackwell, Mar. 2009.

[20] X. Wang and H. Yu, "How to break MD5 and other hash functions," IN EUROCRYPT, vol. 3494, 2005.

[21] P. Hawkes, M. Paddon, and G. G. Rose, "Musings on the wang et al. md5 collision," Cryptology ePrint Archive, Report 2004/64, 2004.

[22] D. Eastlake 3rd and P. Jones, "US Secure Hash Algorithm 1 (SHA1)," RFC 3174 (Informational), Internet Engineering Task Force, Sep. 2001, updated by RFC 4634. [Online]. Available: http://www.ietf.org/rfc/rfc3174.txt [Accessed: 1. Nov 2011]

[23] "Twisted Matrix Labs." [Online]. Available: http://twistedmatrix.com [Accessed: 1. Nov 2011]

[24] J. Undery, "Ieft draft: SIP authentication: SIP digest access authentication," IETF, Tech. Rep., Jul. 2001.

[25] F. Palmieri, "Improving authentication in voice over IP infrastructures," in Advances in Computer, Information, and Systems Sciences, and Engineering, K. Elleithy, T. Sobh, A. Mahmood, M. Iskander, and M. Karim, Eds. Springer Netherlands, 2006, pp. 289 – 296. [Online]. Available: http://www.springerlink.com/content/pj11582775h177q0/ [Accessed: 1. Nov 2011]

[26] F. Palmieri and U. Fiore, "Providing true end-to-end security in converged voice over IP infrastructures," Computers & Security, vol. 28, no. 6, pp. 433–449, Sep. 2009.

[27] C. Yang, R. Wang, and W. Liu, "Secure authentication scheme for session initiation protocol," Computers & Security, vol. 24, no. 5, pp. 381–386, Aug. 2005.

[28] International Telecommunication Union, "H.323 security: Framework for security in H-series (H.323 and other H.245-based) multimedia systems," ITU-T Recommendation H.235.0, 2005.

[29] ——, "H.323 security: Framework for secure authentication in RAS using weak shared secrets," ITU-T Recommendation H.235.5, 2005.

[30] M. Spencer, B. Capouch, E. Guy, F. Miller, and K. Shumard, "IAX: Inter-Asterisk eXchange Version 2," RFC 5456 (Informational), Internet Engineering Task Force, Feb. 2010. [Online]. Available: http://www.ietf.org/rfc/rfc5456.txt [Accessed: 1. Nov 2011]

[31] R. Rivest, "The MD5 Message-Digest Algorithm," RFC 1321 (Informational), Internet Engineering Task Force, Apr. 1992. [Online]. Available: http://www.ietf.org/rfc/rfc1321.txt [Accessed: 1. Nov 2011]

[32] J. Peterson, "S/MIME Advanced Encryption Standard (AES) Requirement for the Session Initiation Protocol (SIP)," RFC 3853 (Proposed Standard), Internet Engineering Task Force, Jul. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3853.txt [Accessed: 1. Nov 2011]

[33] C. Jennings, J. Peterson, and M. Watson, "Private Extensions to the Session Initiation Protocol (SIP) for Asserted Identity within Trusted Networks," RFC 3325 (Informational), Internet Engineering Task Force, Nov. 2002, updated by RFC 5876. [Online]. Available: http://www.ietf.org/rfc/rfc3325.txt [Accessed: 1. Nov 2011]

[34] J. Peterson and C. Jennings, "Enhancements for Authenticated Identity Management in the Session Initiation Protocol (SIP)," RFC 4474 (Proposed Standard), Internet Engineering Task Force, Aug. 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4474.txt [Accessed: 1. Nov 2011]

[35] J. Linn, "Generic Security Service Application Program Interface Version 2, Update 1," RFC 2743 (Proposed Standard), Internet Engineering Task Force, Jan. 2000, updated by RFC 5554. [Online]. Available: http://www.ietf.org/rfc/rfc2743.txt [Accessed: 1. Nov 2011]

 ISBN:978-1-61208-113-7

Performance Evaluation of Inter-Vehicle Communications Based on the Proposed IEEE 802.11p Physical and MAC Layers Specifications

Diogo Acatauassu, Igor Couto, Patrick Alves
Signal Processing Laboratory (LaPS)
Federal University of Pará (UFPA)
Brazil
Email: {diogoaca, icouto, patrickalves}@ufpa.br

Kelvin Dias
Center of Informatics
Federal University of Pernambuco (UFPE)
Brazil
Email: kld@cin.ufpe.br

***Abstract*—Traffic control, accidents prevention, vehicles automation and useful services to the drivers have always been goals of an intelligent traffic system. With this objective, the IEEE is finalizing its new standard: the IEEE 802.11p, which defines the vehicular ad-hoc networks physical and medium access control layers characteristics. This paper presents simulations to evaluate the performance of these networks operating according to the new standard, at different scenarios, using the most recent version of the well-known network simulator NS-2. The results show that the transmissions quality impact is directly linked to dynamic changes in the network topology.**

***Keywords*—*IEEE 802.11p*; *Vehicular Networks*; *NS-2.34*.**

I. Introduction

Big cities all over the world are suffering, or can suffer in the near future, with the uncontrolled growth of their road systems, making the search for solutions that lead to this improvement becomes a challenge [1], [2], [3]. Linked to this, there is a lack of traffic information to drivers, what prevents them to make decisions that could avoid traffic jams. In this scenario, the concept of intelligent transportation system (ITS) was created, where essential data are exchanged by the vehicles, such as: track and weather conditions, levels of traffic jams and accidents emergency announcements. These and other measures would be relevant to the planning of routes and safety of drivers and pedestrians [3], [4], [5], [6].

From this situation came the necessity for the criation of vehicular ad-hoc networks (VANETS), which are able to supply the demand for inter-vehicle communications (IVC) and road-to-vehicle communications (RVC). Such technology is getting lots of attention by both the automotive industry and world research centers [4], [7].

To VANETS's standardization, the IEEE is finalizing a new standard: the IEEE 802.11p, which defines the rules for wireless access in vehicular environment (WAVE) [3], [8], [9], [10]. The new model comes as an alternative to the currents wi-fi standards, being developed to support the vehicular networks features, where the main difficulty is keeping the transmission rates due to the network topology dynamism and nodes high speed, besides low latency in security applications [3], [6].

As this proposed standard is not finished, computer simulations to evaluate its performance are very important to both researchers and industry, being the focus of this work. With this target, experiments were performed using the network simulator NS-2.34, applying the IEEE 802.11p support, where it was observed two main performance parameters of VANETS: packets delay and data throughput.

Some related works can be highlighted, such as [11], [12], [13]; however, in all of this works, the simulations was performed using old NS-2 versions and different VANETS implementations, developed by each one of these authors. This work uses the newer NS-2 version and its native VANETS modules, developed by [14], being different from the latter by the analyzed metrics: packets delay and data throughput.

The paper is organized as follows. First, Section II describes the IEEE 802.11p physical and medium access control layers characteristics. Continuing, Section III describes the IEEE 802.11p implementation in NS-2.34. Then, Section IV presents experiments and results of vehicular networks simulations, using NS-2.34, at different scenarios. Finally, Section V shows the work final considerations and conclusions.

II. IEEE 802.11p Standard

Due to studies, none of the currents wireless standards are completely adapted to VANETS [6]. So, the IEEE is developing a new standard in order to follow vehicle networks requirements with safety and quality, ensuring data transmission in unstable networks. The IEEE 802.11p, with its drafts, defines the operation mode settings of VANETS's physical and medium access control (MAC) layers [3], [8], [9], [10].

The goal of this new proposal is to ensure robust and quality communications when dealing with networks whose nodes have high mobility and fast topology changes, beyond the necessity of low latency and immunity to interference. The IEEE 802.11p definitions are described in Subsections II-A and II-B.

 ISBN:978-1-61208-113-7

TABLE I
IEEE 802.11P AND IEEE 802.11A MAIN PARAMETERS COMPARISON

Parameter	IEEE 802.11p	IEEE 802.11a
Rate (Mbps)	3, 4.5, 6, 9 12, 18, 24 and 27	6, 9, 12, 18 14, 36, 48 and 54
Modulation	BPSK, QPSK 16-QAM and 64-QAM	BPSK, QPSK 16-QAM and 64-QAM
Codification Rate	1/2, 1/3 and 3/4	1/2, 1/3 and 3/4
Sub-carriers Number	52	52
OFDM Symbol Duration	8μs	4μs
Guard Interval	1.6μs	0.8μs
FFT Period	6.4μs	3.2μs
Preamble Duration	32μs	16μs
Sub-carriers Spacing	0.15625 MHz	0.3125 MHz

TABLE II
CONTROL CHANNEL EDCA PARAMETERS.

AC	CWmin	CWmax	AIFS
0	CWmin	CWmax	9
1	(CWmin+1)/2-1	CWmin	6
2	(CWmin+1)/4-1	(CWmin+1)/2-1	3
3	(CWmin+1)/4-1	(CWmin+1)/2-1	2

TABLE III
SERVICE CHANNELS EDCA PARAMETERS.

AC	CWmin	CWmax	AIFS
0	CWmin	CWmax	7
1	CWmin	CWmax	3
2	(CWmin+1)/2-1	CWmin	2
3	(CWmin+1)/4-1	(CWmin+1)/2-1	2

A. Physical Layer

The IEEE 802.11p physical layer implementation specifies the use of dedicated short range communications (DSRC), defined by the Federal Communications Commission (FCC) [11].

The DSRC technology operates at a 75 MHz bandwidth, positioned in the spectrum range of 5.9 GHz. These 75 MHz are divided in seven 10 MHZ channels each, being the center channel the control channel and the rest of the channels the service channels [4], [11], [12], as illustrated in Fig. 1.

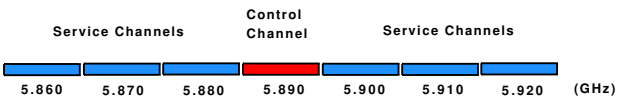


Figure 1. DSRC spectrum.

Different channels can not be used simultaneously, thus, each station can make the constant change between the control channel and the service channels. To ensure the requirement of low delay, especially when safety data are sent, the changing time can not be higher than 100 *m*s [4].

During transmissions, signals are sent using orthogonal frequency division multiplexing (OFDM) technique, which divides each channel in several sub-carriers spaced by 0.15625 MHz from each other [3].

To illustrate the differences, Table I compares the IEEE 802.11p and IEEE 802.11a physical layer main parameters.

B. MAC Layer

The MAC layer functions match to the IEEE 802.11e standard, enhanced distributed channel access (EDCA), which adds quality of service to IEEE 802.11 networks. Messages are categorized into four different ACs (AC0, AC1, AC2 and AC3), where AC0 has the lowest priority and AC3 has the highest priority [10].

When a particular message is selected, its contain parameters are sent to the transmitter. First the arbitrary inter frame space (AIFS), previously set for each AC. As each slot time is 16 μs, the AIFS time is equal to AIFS x 16 μs. Subsequently is calculated the contention window (CW) time. This is performed by a random value between 0 and CWmin. If there is any collision, the window time is recalculated by $2(CW + 1) - 1$, and a new attempt is done. The operation is repeated until the maximum window size (CWmax) is reached or the packet is sent successfully [4], [10].

The control and service channels contain parameters are shown in Tables II and III, respectively.

III. VANETS SIMULATION SUPPORT

VANETS performance evaluation is being studied by several researchers, [4], [5], [11], [12], and are very important for automotive industry. Although testbeds are still limited, due to the fact that IEEE 802.11p is not finished, computer simulations can be performed and its results used as a parameter for possible vehicular networks improvements. In this scenario we can highlight the use of NS-2.

NS-2 is a general purpose networks simulator developed by Berkley University [15] and is currently at version 2.34. VANETS support, however, was only developed in the last two versions (2.33 and 2.34).

Its implementation is done by applying the IEEE 802.11p physical and MAC layers features in the TCL simulation code, defined by two native modules: *WirelessPhyExt* and *MAC80211-Ext*.

Table IV shows the definitions of some vehicular networks key parameters in NS-2 TCL simulation code.

IV. EXPERIMENTS AND RESULTS

Using NS-2.34 VANETS's support, performance evaluation experiments were realized.

However, it was necessary to check whether this implementations was sufficient to obtain consistent results because, as described in Section III, the NS-2 IEEE 802.11p modules were recently developed, being found only in the two latest versions of the simulator.

So, foremost, an IEEE 802.11a and IEEE 802.1p comparison scenario was simulated, where data throughput and packets delay were verified. These experimental results, if consistent, would enable a more secure analysis in a scenario implemented only using IEEE 802.11p for VANETS simulations.

 ISBN:978-1-61208-113-7

TABLE IV
IEEE 802.11P PARAMETERS DEFINITIONS IN NS-2.34 TCL CODE

Phy/WirelessExt set Pt 5.0ϵ-2 *Phy/WirelessExt set freq 5.85ϵ+9* *Phy/WirelessExt set HeaderDuration 0.000040* *Phy/WirelessExt set BasicModulationScheme 0* *Phy/WirelessExt set PreambleCaptureSwitch 1* *Phy/WirelessExt set DataCaptureSwitchSwitch 0* *Phy/WirelessExt set SINRPreambleCapture 2.5118* *Phy/WirelessExt set SINRDataCapture 100.0* *Phy/WirelessExt set PHY-DBG 0* *Phy/WirelessExt set bandwith 70ϵ6*
MAC/80211Ext set CWmin 15 *MAC/80211Ext set CWmax 1023* *MAC/80211Ext set SlotTime 0.000016* *MAC/80211Ext set SIFS 0.000032* *MAC/80211Ext set ShortRetryLimit 7* *MAC/80211Ext set LongRetryLimit 4* *MAC/80211Ext set HeaderDuration 0.000040* *MAC/80211Ext set SymbolDuration 0.000008* *MAC/80211Ext set BasicModulationScheme 0* *MAC/80211Ext set use80211aFlag true* *MAC/80211Ext set RTSThreshold 2346* *MAC/80211Ext set MACDBG 0*

Then, a simple scenario containing two nodes (transmitter and receiver) was defined, with 100 m x 3000 m topology at a 10 seconds simulation. The nodes movement was done in opposite directions and in each simulation the speed of each node was increased by 20 km/h. Fig. 2 illustrates the proposed scenario.

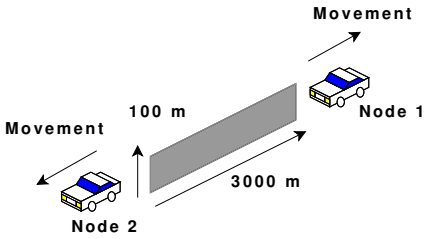


Figure 2. IEEE 802.11p and IEEE 802.11a comparison scenario.

Starting at an initial speed of 40 km/h and increased by 20 km/h until the final speed of 140 km/h, it was possible to obtain the packets delay and the transmitted data throughput. The results are shown in Fig. 3 and Fig. 4, respectively.

The graphs show clearly the best performance of IEEE 802.11p implementation compared to IEEE 802.11a.

Due to the increased power at the transmitter, [3], [8], [9], [10], and being exposed to the same propagation model (in this case the Nakagami model [13]), the IEEE 802.11p achieved considerably greater data throughput, especially at speeds below 80 km/h. Analyzing the delay, the NS-2 IEEE 802.11p implementation proved to be robust and promoted the low latency support required in the standard, [3], [8], [9], [10], resulting in average delays almost 3 times smaller than those ones obtained using IEEE 802.11a.

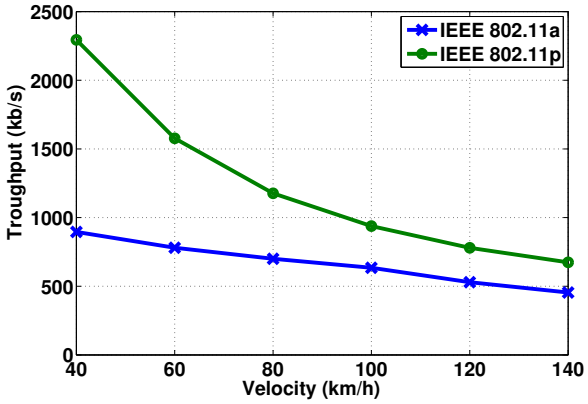


Figure 3. IEEE 802.11p and IEEE 802.11a data throughput comparison.

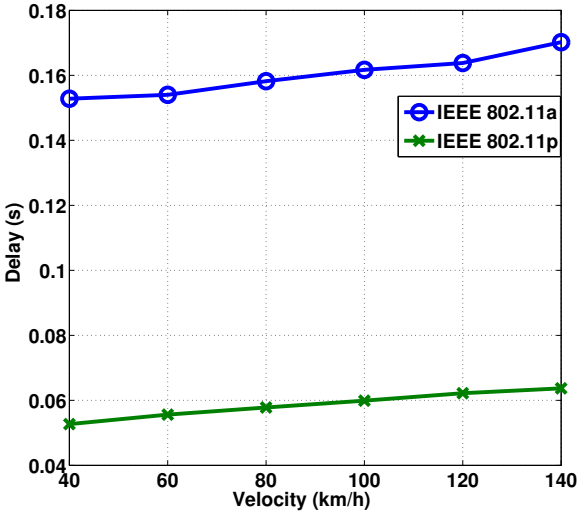


Figure 4. IEEE 802.11p and IEEE 802.11a packets delay comparison.

With these results, which showed the effectiveness of the NS-2.34 IEEE 802.11p implementation, a new series of experiments was performed. In these tests were analyzed vehicular networks performance at different scenarios using the same two metrics: throughput and delay.

The scenarios were defined according to three variables: number of nodes (10, 30 or 50 nodes), nodes average speed (70 km/h, 90 km/h and 110 km/h) and nodes average distance (10 m, 30 m or 50 m), totalizing 27 different scenarios.

The experiments were performed by TCP transmissions realized by two nodes located at the opposite sides of a road with 100 m x 3000 m topology, using AODV routing protocol, Nakagami propagation model (as this is the more accurate to characterize vehicular networks communications [13]), and the IEEE 802.11p physical and MAC layers implementations, in 10 seconds simulations. Table V shows the parameters definition in the TCL code used during the simulations, and Fig. 5 illustrates the scenarios general arrangement.

In the first experiment was observed if increasing the nodes

 ISBN:978-1-61208-113-7

TABLE V
TCL CODE PARAMETERS FOR VANETS PERFORMANCE EVALUATION IN NS-2.34.

```
set val(chan) Channel/WirelessChannel
set val(prop) Propagation/Nakagami
set val(netif) Phy/WirelessPhyExt
set val(mac) Mac/80211Ext
set val(rp) AODV
set val(x) 100
set val(y) 3000
set val(stop) 10.0
```

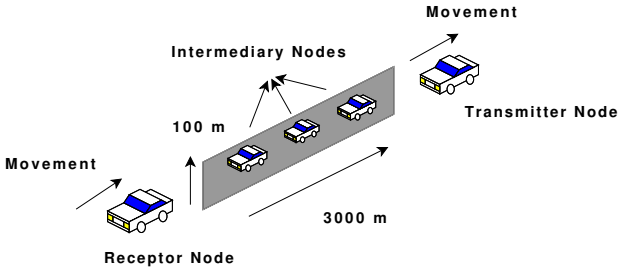


Figure 5. VANETS performance evaluation scenario

average speed (70 km/h, 90 km/h and 110 km/h), fixing the nodes average distance (10 m, 30 m and 50 m) could cause an impact on the data throughput. Fig. 6 illustrates the result, where n is the number of nodes and d is the average distance between them.

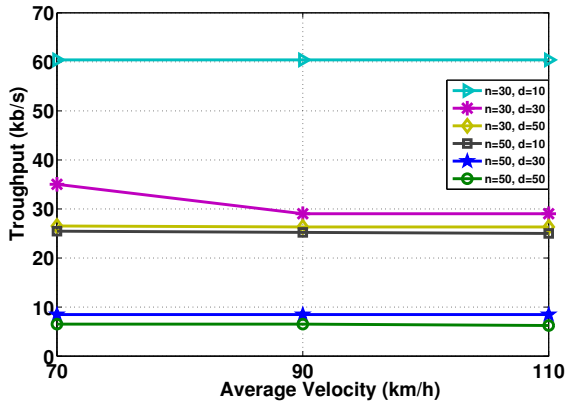


Figure 6. Data throughput changing nodes average velocity.

The results showed that the data throughput is almost invariant when the nodes average velocity remains relatively constant in IVC transmissions. However, the nodes distance influences the data throughput, being inversely proportional to the same. Thus one can imagine the standard is well-established for data transmissions on highways that allows this type of situation, as the German Autobhans for example [16], where cars can run at high speeds, forming blocks according to the adopted velocity.

For urban scenarios transmissions, where nodes average distance and average speed is constantly changing, although not simulated, the standard suggests the use of fixed infrastructure, applying RVC transmissions, in order to keep the data throughput stable in most of the cases [3].

The second experiment used the same procedure as the first, this time analyzing the packets delay. The result is illustrated in Fig. 7, where n is the number of nodes and d is the average distance between them.

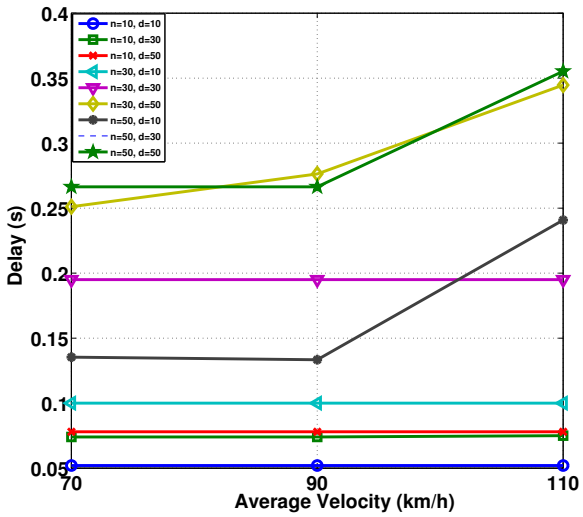


Figure 7. Packets delay changing nodes average velocity.

In this case, although some outliers, mainly for speeds above 90 km/h, the results showed that in most of the cases the delay was constant, confirming the first experiment results.

In the third experiment, the average speed of the nodes was fixed at 70 km/h; the impact on data throughput was verified by changing the number of nodes and the average distance between them. The result is illustrated in Fig. 8, where n is the number of nodes.

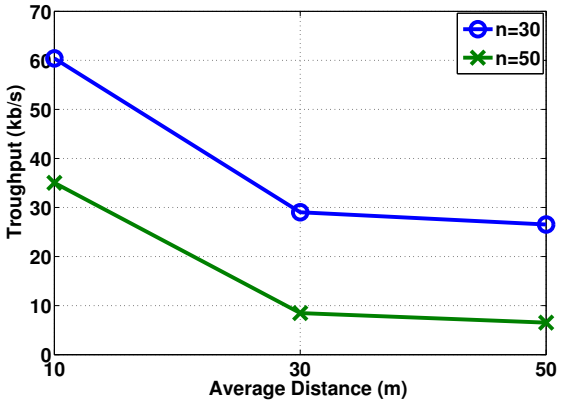


Figure 8. Data throughput for a fixed velocity of 70 km/h changing the number of nodes and the average distance between them.

The results show that for a constant velocity the data throughput impact is given by the network topology. In this

case the data throughput rate in a transmission of two nodes located in the opposite points of a road, with 30 intermediate routing nodes between them, can reach 5 percent of the value obtained in a two nodes direct transmission, as shown in the results of Fig. 3.

Finally, to prove that the network topology change has a greater impact on VANETS performance, was observed the packets delay of a simulation where the nodes average speed was fixed in 90 km/h and was changed the number of nodes and the average distance between them. Fig. 9 illustrates the result, where *n* is the number of nodes.

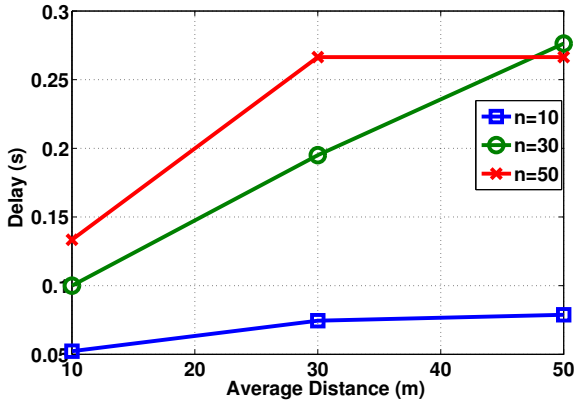


Figure 9. Packets delay for a fixed velocity of 90 km/h changing the number of nodes and the average distance between them.

The results show that the impact of changing the network topology is also noticeable in packets delay.

V. Conclusion

Due to big cities road systems growth, a lot of attention is directed to the development and performance of vehicular networks, not only by manufacturers and researchers, but also by governments and institutions responsible for maintenance of these roads.

While IEEE does not finish the IEEE 802.11p standard, which defines the VANETS physical layer and MAC layer characteristics, real testbeds are still limited.

Therefore, computational experiments become the most widely used tool for obtaining these performance parameters, and their results can be used for possible changes in the new standard specifications.

This paper described computer based simulations, using NS-2.34, whose goal was to obtain the performance of two major optimization parameters in VANETS, packets delay and data throughput, for vehicular networks that implements direct communication between the nodes, IVC.

The results showed that increasing vehicles average speeds and keeping the average distance constant, for a given number of nodes, the impact on delay and throughput was low. That is, the standard provides good support for vehicles communications in scenarios such as highways, for example.

Furthermore, it was found that keeping the nodes average speed, the impact on data throughput and packets delay is given directly by the network topology, that is the number of nodes and average distance between them, being a possible limiting factor for the VANETS performance which adopts only IVC transmissions.

Acknowledgement

The authors gratefully acknowledge CNPq for financial support.

References

[1] L. Vita and M. C. Marolda, “Road infrastructure - the backbone of transport system,” EUROPEAN COMMISSION, Tech. Rep., 2008.

[2] N. M. Rabadi and S. M. Mahmud, “Performance evaluation of IEEE 802.11a MAC protocol for vehicle intersection collision avoidance system,” in *4th IEEE Consumer Communications and Networking Conference, 2007. CCNC 2007.*, jan 2007, pp. 54–58.

[3] “IEEE trial-use standard for wireless access in vehicular environments (WAVE)–networking services,” *IEEE Std 1609.3-2007*, pp. c1–87, 20 2007.

[4] S. Eichler, “Performance evaluation of the IEEE 802.11p WAVE communication standard,” in *IEEE 66th Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007.*, 30 2007-Oct. 3 2007, pp. 2199–2203.

[5] J. Blum, A. Eskandarian, and L. Hoffman, “Challenges of intervehicle ad hoc networks,” *IEEE Transactions on Intelligent Transportation Systems.*, vol. 5, no. 4, pp. 347–351, Dec. 2004.

[6] H. Oh, C. Yae, D. Ahn, and H. Cho, “5.8 GHz DSRC packet communication system for ITS services,” in *IEEE VTS 50th Vehicular Technology Conference, 1999. VTC 1999 - Fall.*, vol. 4, 1999, pp. 2223–2227 vol.4.

[7] F. Karnadi, Z. H. Mo, and K. chan Lan, “Rapid generation of realistic mobility models for VANET,” in *IEEE Wireless Communications and Networking Conference, 2007.WCNC 2007.*, March 2007, pp. 2506–2511.

[8] “IEEE trial-use standard for wireless access in vehicular environments - security services for applications and management messages,” *IEEE Std 1609.2-2006*, pp. c1–105, 2006.

[9] “IEEE trial-use standard for wireless access in vehicular environments (WAVE)–networking services,” *IEEE Std 1609.3-2007*, pp. c1–87, 20 2007.

[10] “IEEE trial-use standard for wireless access in vehicular environments (WAVE) - multi-channel operation,” *IEEE Std 1609.4-2006*, pp. c1–74, 2006.

[11] B. Gukhool and S. Cherkaoui, “IEEE 802.11p modeling in NS-2,” in *33rd IEEE Conference on Local Computer Networks, 2008. LCN 2008.*, Oct. 2008, pp. 622–626.

[12] T. Murray, M. Cojocari, and H. Fu, “Measuring the performance of IEEE 802.11p using NS-2 simulator for vehicular networks,” in *IEEE International Conference on Electro/Information Technology, 2008. EIT 2008.*, May 2008, pp. 498–503.

[13] F. Schmidt-Eisenlohr, M. Torrent-Moreno, J. Mittag, and H. Hartenstein, “Simulation platform for inter-vehicle communications and analysis of periodic information exchange,” in *Fourth Annual Conference on Wireless on Demand Network Systems and Services, 2007. WONS ’07.*, Jan. 2007, pp. 50–58.

[14] 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,” in *MSWiM ’07: Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*. New York, NY, USA: ACM, 2007, pp. 159–168.

[15] Y. Xue, H. S. Lee, M. Yang, P. Kumarawadu, H. Ghenniwa, and W. Shen, “Performance evaluation of NS-2 simulator for wireless sensor networks,” in *Canadian Conference on Electrical and Computer Engineering, 2007. CCECE 2007.*, April 2007, pp. 1372–1375.

[16] C.-H. Rokitansky and C. Wietfeld, “Methods and tools for performance evaluation and validation of vehicle-roadside communications proposed for standardization,” in *IEEE 45th Vehicular Technology Conference, 1995*, vol. 2, Jul 1995, pp. 964–970 vol.2.

 ISBN:978-1-61208-113-7

Cooperative Vehicle Information Delivery Scheme for ITS Networks with OFDM Modulation Techniques

Katsuhiro Naito, Kazuo Mori, and Hideo Kobayashi
Department of Electrical and Electronic Engineering, Mie University,
1577 Kurimamachiya, Tsu, 514-8507, Japan
Email: {naito, kmori, koba}@elec.mie-u.ac.jp

Abstract— **Vehicular Ad Hoc Networks (VANETs) are new technologies that offer many opportunities to wide range interesting services. Safe driving applications in Intelligent Transport System (ITS) are major applications of VANETs. In the high-speed mobility environment of VANETs, failure transmission due to change of vehicle positions, fluctuation of channel condition and blocking by large vehicles may decrease reachability of vehicle information messages for safety usage. In this paper, we focus on an OFDM transmission technology, which is employed in IEEE 802.11p for ITS networks. In the proposed scheme, some vehicles forward a same OFDM signal at almost same instance, which is less than a guard interval period. Therefore, vehicles can demodulate some same OFDM signals from different vehicles, and can obtain path diversity effect through some different vehicles. This paper also proposes a new media access control scheme to achieve the proposed transmission technology. The proposed media access control scheme is based on carrier sense multiple access (CSMA). Meanwhile, some forwarder vehicles can select the same random back-off period autonomously to synchronize transmission timing. As the results, the proposed scheme can achieve the high delivery ratio of vehicle information messages and reduce transmission delay. Finally, we consider vehicle movements, channel fluctuation due to fading and blocking due to large-size vehicle in computer simulations. The numerical results show that the proposed scheme can achieve the high delivery ratio with the short delay even if actually real environment, which considers fast movement, blocking, fading, is evaluated in the simulations. Moreover, we clarify that our scheme has high scalability in case of increasing of vehicles.**

Keywords— **VANET, ITS networks, OFDM, Media access control, Vehicle information**

I. INTRODUCTION

Vehicular ad-hoc networks (VANETs) have been focused recently for building Intelligent Transportation Systems (ITS) [1], [2]. In VANETs, many vehicles construct a temporal network autonomously to communicate each other. VANETs have special attributes that differentiate it from the other types of networks such as mobile ad hoc networks (MANETs) [3], [4], [5]. Especially, the mobility patterns of vehicles in VANETs are more restrictive due to road structures. Therefore, almost all vehicles can only communicate with front and backward vehicles. In the conventional works, many routing protocols have been proposed to achieve effective data dissemination. These protocols are classified into some categories such as pure ad-hoc routing, position-based routing, and broadcast routing.

In MANETs, several routing protocols have been proposed and are still applicable for VANETs. Ad-hoc On-demand Distance Vector (AODV) [6] and Dynamic Source Routing (DSR) [7] are well-known routing protocols for general purpose mobile ad-hoc networks. Meanwhile, VANETs differ from MANETs by their dynamic change of network topology. In conventional studies, most pure ad-hoc routing protocols suffer from highly dynamic nature of vehicle mobility and tend to have low communication throughput due to poor route management performance [8].

Position-based routing employs routing strategies that use geographical information obtained from on-board navigation systems because movement of vehicles is restricted in just bidirectional movements constrained along roads and streets [9]. Most position-based routing algorithms are based on forwarding decision upon location information [10], [11], [12]. Additionally, some protocols consider the connectivity to construct reliable routes [13], [14]

Broadcast routing is frequently used for delivering advertisements and announcements in VANETs [15]. The simplest way is flooding, in which each vehicle re-broadcasts packets to all of its neighbors. Flooding performs relatively well for a small number of vehicles [16], [17]. However, it suffers from broadcast storm problems when the number of vehicle in networks increases because a lot of redundant messages are re-broadcasted and many collisions occur in networks [18]. Some schemes for the broadcast storm problems have been proposed in ad hoc networks [19], [20]. However, the investigation about the broadcast storm problems is not enough to be considered in VANETs [21].

Meanwhile, the IEEE 802.11p, intended for vehicular communication, has drown attention recently [22], [23]. The IEEE 802.11p also employs carrier sense multiple access (CSMA) mechanisms as medium access control techniques. Therefore, vehicles first listen to channel and transmit data packets if the channel has been free for a certain period. Hence, several transmissions are performed when we employ broadcast

 ISBN:978-1-61208-113-7

routing, and these transmissions cause long delay and packet collisions.

The physical layer of the IEEE 802.11p employs Orthogonal Frequency Division Multiplexing (OFDM) as modulation techniques. The OFDM has been focused for high-speed data transmission in wireless LAN, cellular systems and etc. OFDM signals are multipath robust due to low symbol rates and addition of a guard interval (GI) to an OFDM symbol [24]. Therefore, multipath reflections that have a delay spread less than the guard interval period can be demodulated with no inter-symbol interference (ISI). This characteristic is used for Single Frequency Networks (SFN) in television broadcast systems [25]. In SFN, the same OFDM signal is transmitted from some fixed antenna towers, which exist in different places. In order to demodulate multiple OFDM signals without ISI, reception timing of these OFDM signals should be less than the guard interval period. Therefore, high accuracy transmission timing control is required at transmitters.

Authors consider that the concepts of SFN can apply to vehicular networks. Meanwhile, vehicular networks have some different characteristics from television broadcast systems. First different characteristic is fast movement of vehicles. Therefore, physical relationship between vehicles is always changing, and forwarding vehicles are almost changing according to the physical locations. Therefore, autonomous transmission timing control mechanisms are required to achieve the concepts of SFN in vehicular networks. If we can achieve this concept, vehicles can obtain path diversity effect. As the results, communication performance will be improved without additional wireless resource.

In this paper, we focus on cooperative multiple transmission schemes similar to SFN for vehicular networks to improve transmission performance and to reduce transmission delay. Then, we propose a new vehicular network with cooperative transmission mechanisms. In the proposed vehicular network, neighbor vehicles select the same random delay period for collision avoidance autonomously, and forward the same OFDM signal at same instance according to the selected random delay period. Vehicles can demodulate the received OFDM signals without ISI when the arrival timing of each OFDM signals is confined to the guard interval period of the OFDM signals. Moreover, we assume the different sizes of vehicles in the computer simulations. Then, we evaluate the proposed scheme in the more actual wireless environment. The numerical results show that the proposed scheme can achieve the high delivery ratio with the short delivery delay.

II. System Model

In this paper, we intend to achieve data dissemination of vehicle information messages in specific area near a vehicle. Then, we employ broadcast communication to simplify transmission mechanisms to reduce end-to-end delay. In the almost all studies of ad-hoc networks, only one signal is transmitted by only one node. Therefore, more wireless resource is required to achieve path diversity effects because some nodes transmit the same signal at different timing. As the result, it is well known that broadcast storm problems occur.

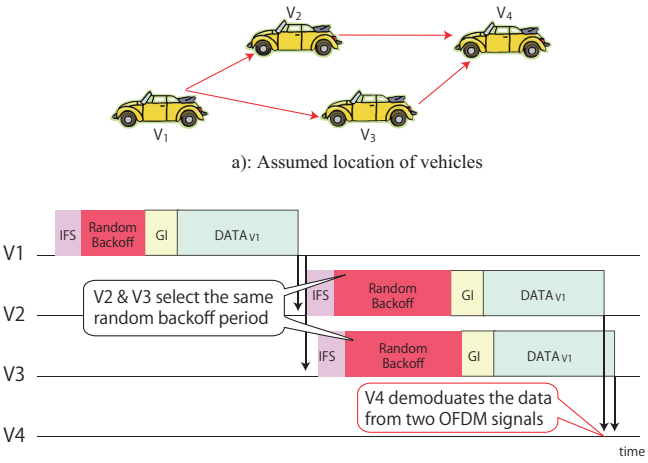


Fig. 1. Overview communication.

Meanwhile, some vehicles transmit the same OFDM signal simultaneously in the proposed scheme. Therefore, transmission period of the proposed scheme equals to that of the single signal transmission. Hence, our scheme can achieve path diversity effects without additional consumption of wireless resource. Additionally, vehicles re-broadcast the received packets of vehicle information messages when the vehicles exist in delivery area of the source vehicle that transmits the received packets.

Figure 1 shows the overview communication of the proposed scheme. Figure 1 a) shows the assumed location of four vehicles, and Fig. 1 b) shows the packet transmission timing. In the assumptions, the vehicle V_1 transmits its own vehicle information message to neighbor vehicles V_2 and V_3 by broadcasting. In order to demodulate two signals without inter-symbol interference at the vehicle V_4, the two signals from the vehicles V_2 and V_3 must arrive during the guard interval period. Hence, the vehicles V_2 and V_3 should select the same random back-off value. Then, these two vehicles forward the same OFDM signal at almost same instance. As the results, the vehicle V_4 receives the two same signals from V_2 and V_3, and obtain the path diversity effect.

It is known that wireless channel in vehicular networks is assumed to be fading environments. In fading environments, only one wireless link may not be enough to achieve reliable communication. Additionally, some broadcast based protocols have been considered to achieve safety applications, which im-

 ISBN:978-1-61208-113-7

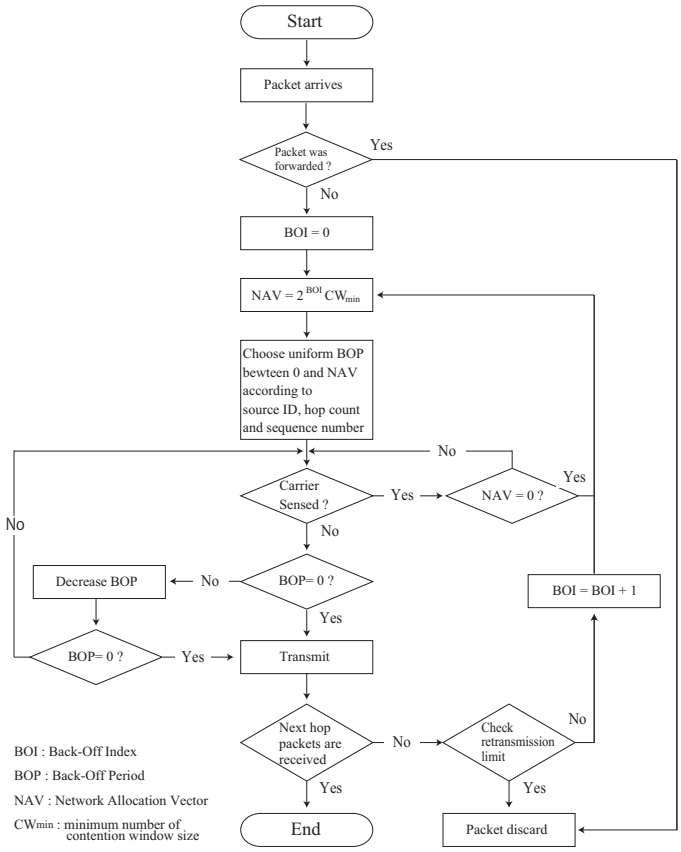
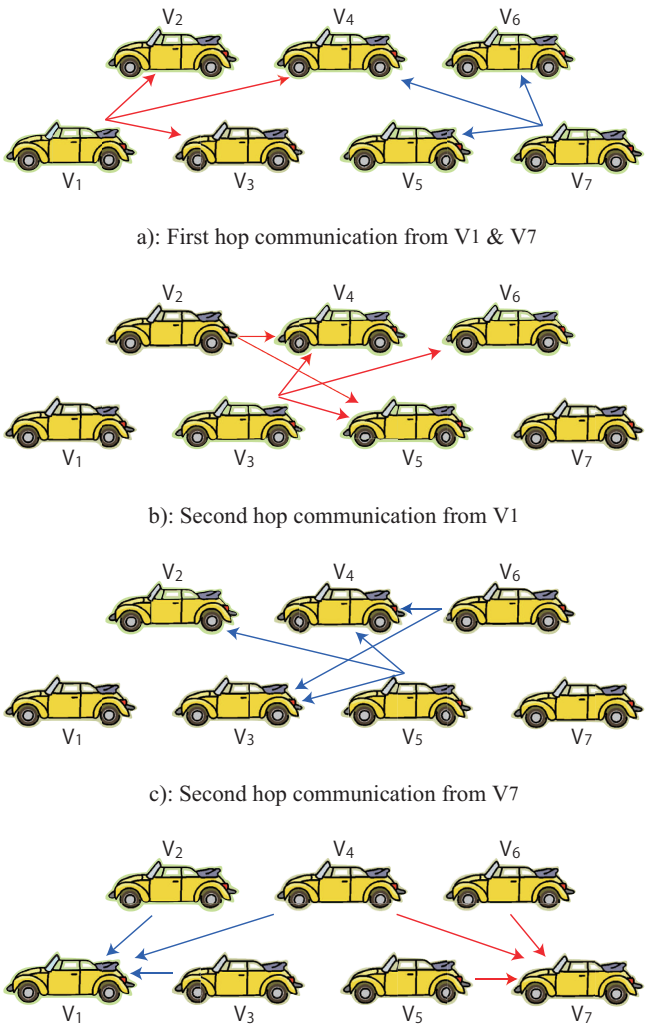


Fig. 2. Flowchart of media access control.



a): First hop communication from V1 & V7

b): Second hop communication from V1

c): Second hop communication from V7

d): Third hop communication from V1 & V7

Fig. 3. Example of transmission procedures.

prove vehicle safety on the roads. Meanwhile, the transmission period of the proposed scheme equals to that of conventional CSMA mechanisms. Therefore, the proposed scheme can reduce transmission delay with performance improvement.

The proposed scheme employs the source ID, the hop count information and the sequence number of data packets for controlling transmission timing. Therefore, prior negotiation process is not required to start forwarding of data packets. Then, the proposed scheme can also be employed for collision avoidance systems in crossroads.

Figure 2 shows the flowchart of the proposed media access control scheme. The proposed scheme is extended mechanisms of CSMA/CA. In general CSMA mechanisms, random back-off periods are selected randomly at each node. Therefore, nodes have different random back-off periods, and transmit a packet at different timing to avoid packet collisions. In the proposed scheme, vehicles that transmit the same OFDM signal have to select the same random back-off period autonomously. Moreover, we assume that the proposed scheme should be processed in preference to schemes for single path transmission because interruption due to other transmission of packets damages performance of the proposed scheme. The procedures of the proposed scheme are described as follows.

- The procedures start when vehicles receive a packet from neighbor vehicles.
- Vehicles confirm the packet forwarding history to avoid redundant forwarding.
- Vehicles initialize the Back-Off Index (BOI) that determines the Network Allocation Vector (NAV), where CW_{min} is the minimum number of contention window size.
- Vehicles calculate the new NAV according to the initialized BOI value. The NAV denotes the interval period for the back-off period (BOP). In the general CSMA mechanisms, the BOP is selected randomly with uniform distribution.
- Vehicles generate a random value with the source ID, the hop count information and the sequence number of data packets for each node as random seeds. Hence, vehicle

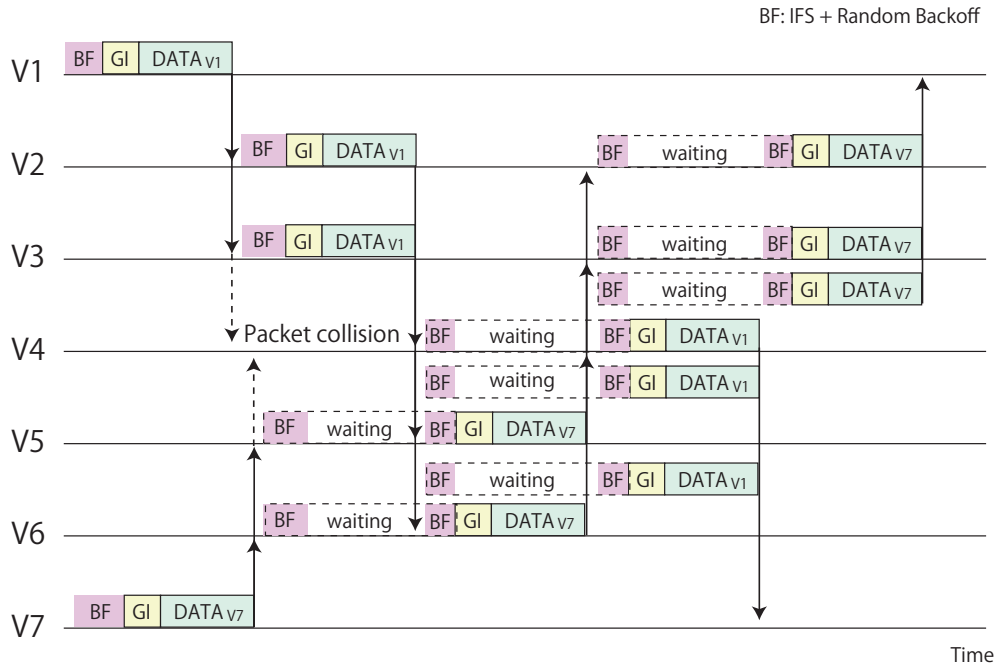


Fig. 4. Example of packet transmission timing.

can obtain the unique random value according to the source ID, the hop count information and the sequence number of data packets. Therefore, some vehicles that forward the same OFDM signal can select the same random back-off periods. As the results, the proposed scheme can synchronize the transmission timing of some forwarding vehicles autonomously. Moreover, vehicles can obtain transmission opportunities randomly because the random back-off period is generated randomly.

- Vehicles sense channel status. If the channel status is busy, vehicles check the NAV value and retry channel sensing. If the channel status is idle, vehicles check the BOP value.
- If the BOP value does not equal to zero, vehicles decrease the BOP value and wait for transmission timing. If the BOP value equals to zero, vehicles start to transmit the OFDM signal.
- After transmission of the OFDM signal, vehicles sense the channel to confirm transmission of the same data packet transmitted by next hop vehicles.
- When vehicles receive the same data packet of the transmitted OFDM signal, the procedures end. If not, vehicles check the retransmission limit of data packets to avoid continuous retransmission. Then, vehicles increase the BOI value to extend the NAV period to reduce collision probability, and try to retransmit the OFDM signal.

III. Operation examples

Figures 3 show the example transmission procedures. Figure 4 shows the packet transmission timing of the proposed schemes. In these figures, both the vehicle V_1 and the vehicle V_7 transmit the packets of their own vehicle information messages to the whole seven vehicles at almost same instance. Each procedure is described as follows.

- Figure 3 a) shows the first hop communication. Therefore, the vehicle V_1 and the vehicle V_7 transmit the packets. Vehicles V_2, V_3 and V_4 receive the signal of the packet from the vehicle V_1. Vehicles V_4, V_5 and V_6 receive the signal of the packet from the vehicle V_7. Therefore, vehicles V_2 and V_3 receive the packet from the vehicle V_1, and vehicles V_5 and V_6 receive the packet from the vehicle V_7. Meanwhile, the vehicle V_4 cannot receive the packets from the vehicle V_1 nor the vehicle V_7 because the two signals from the vehicles V_1 and V_7 conflict at the vehicle V_4.
- Figure 3 b) shows the second hop communication from the vehicle V_1. In this example, the selected random back-off value of the vehicles V_2 and V_3 is assumed to be less than that of the vehicles V_5 and V_6. Therefore, the vehicles V_2 and V_3 transmit the same signal at first. These two signals arrive at the vehicles V_4, V_5 and V_6. Then, these vehicles demodulate the two received signals. Meanwhile, the vehicles V_5 and V_6 resume the back-off process after the transmission from the vehicles V_2 and V_3 is completed.

 ISBN:978-1-61208-113-7

- Figure 3 c) shows the second hop communication from the vehicle V_7. The vehicles V_5 and V_6 transmit the same signal. The vehicles V_2, V_3 and V_4 receive the two signals, and demodulate them.
- Figure 3 d) shows the third hop communication. In this example, the selected random back-off value for the vehicle information message of the vehicle V_1 is assumed to be less than that for the vehicle V_7. Therefore, the vehicles V_4, V_5 and V_6 transmit the same signal from the vehicle V_1 firstly, and the vehicles V_2, V_3 and V_4 also transmit the same signal from the vehicle V_7 secondly. Hence, the vehicle V_1 receives the packet from the vehicle V_7, and the vehicle V_7 receives the packet from the vehicle V_1.

IV. Numerical Results

To evaluate the proposed scheme, we performed computer simulations with network simulator QualNet [26]. Qualnet is the well-known wireless network simulation software that considers the more actual wireless environment. Therefore, the simulator considers packet errors due to low signal-to-interference and noise power ratio (SINR), channel fluctuation due to fading, blocking by large-size vehicles. The results are an average of 10 trial simulations. Our proposed scheme intends to achieve data dissemination of vehicle information messages for safe driving systems. Therefore, we employ broadcast communication to deliver the vehicle information messages in delivery area. In the simulations, the delivery area is set to 1000 [m] from a source vehicle. It is known that broadcast communication suffers from packet collisions when many vehicles exist in a communication area. Therefore, we considered 50 vehicles for small number of vehicles and 300 vehicles for large number of vehicles. We assumed that the road shape is the loop line with the radius equals to 1500 [m] and 2 lanes. Each vehicle is located randomly on the road, selecting the velocity between 90 [km/h] and 110 [km/h] randomly. Therefore, the distribution of vehicle velocity is uniformly between 90 [km/h] and 110 [km/h] The vehicle runs on the inside lane principally and keeps an inter-vehicular distance as 100 [m]. If there is no vehicle on the outside lane, the vehicle moves to the outside lane from the inside lane to overtake a forward vehicle. After overtaking, the vehicle moves to the inside lane if there is no vehicle on the inside lane. In the simulations, the passings occur according to the Table I.

The feature of this paper is also to consider the effect of large-size vehicles. Hence, we define the large-size vehicle ratio (LVR) that means the ratio of the large-size vehicles and the standard-size vehicles. When the large-size vehicle ratio is set to 0, all vehicles are standard-size vehicles. Meanwhile, we assumed that large-size vehicles are rectangular solids.

TABLE I
AVERAGE NUMBER OF PASSINGS.

Number of vehicles	50	100	150	200	250	300
Number of passings	23	27	36	59	185	249

TABLE II
SIMULATION PARAMETERS.

Simulator	QualNet
Simulation time	150 [s]
Simulation trial	10 [times]
Number of vehicles	50 – 300 [vehicles]
Vehicle velocity	90 – 110 [km/h]
Size of vehicle information message	100 [Byes]
Transmission interval	200 [ms]
Communication device	IEEE 802.11p
Transmission rates	6 [Mbps]
Transmission power	19 [dBm]
Channel frequency	5.9 [GHz]
Antenna gain	0 [dB]
Antenna type	Omni directional
Antenna height	1.5 [m]
Propagation path loss model	Free Space
Wireless environment	Rayleigh fading
Road shape	Circle with radius = 1500 [m]
Number of lanes	2 [lanes]

If the rectangular solid is overlapped with the straight line between two standard-size vehicles, these two vehicles cannot communicate due to blocking. Additionally, a free space propagation model is used as the wireless propagation model. To consider channel fluctuation due to movement, rayleigh fading according to 100 [km/h] is assumed in the simulations.

The final purpose of this study is to fuse vehicle information delivery networks and communication networks for several network applications. Therefore, we employ IEEE 802.11p, which is a future communication device for ITS networks. In the simulations, the transmission range is about 285 [m], packet errors are determined due to the received signal-to-interference and noise power ratio (SINR). The size of a vehicle information message is 100 [Byte], and is transmitted with 5 [packets/s].

Our scheme is one of the broadcast communication methods. Therefore, we employ the probabilistic flooding scheme in comparison. The flooding probability is assumed to be 50, 75, and 100 [%]. Meanwhile, retransmission of vehicle information messages is not performed in order to evaluate the proposed scheme and flooding mechanisms fairly. Simulation parameters are shown in detail in Table II.

Figure 5 shows the area delivery ratio of vehicle information messages with the large-size vehicle ratio equals to 0 [%]. In this study, we define that the area delivery ratio is the message received ratio for vehicles in the delivery area. From

 ISBN:978-1-61208-113-7

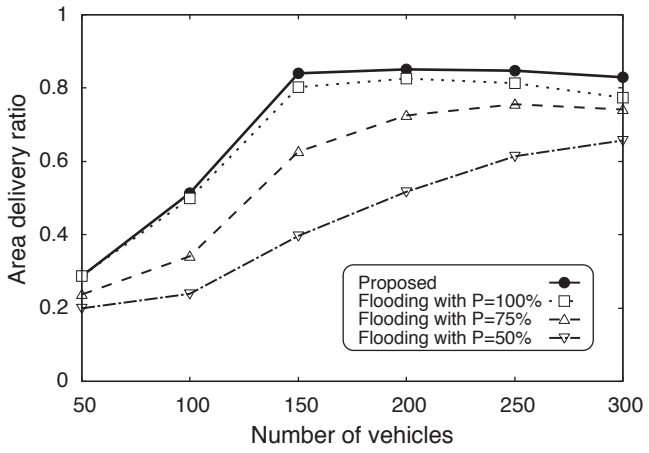


Fig. 5. Area delivery ratio (LVR=0).

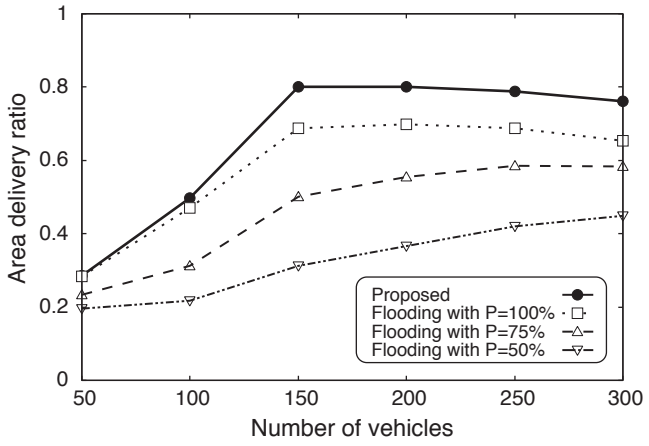


Fig. 6. Area delivery ratio (LVR=0.2).

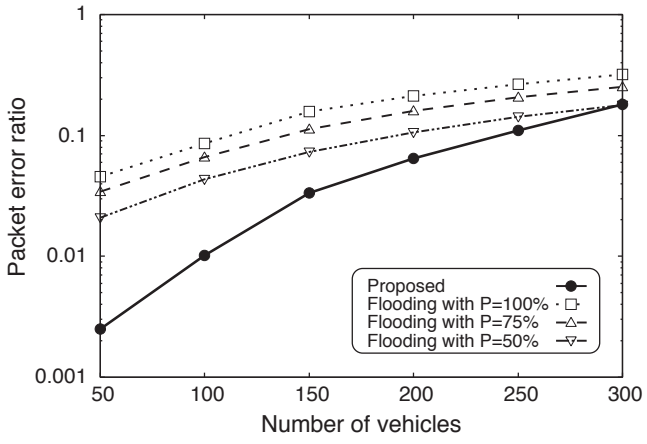


Fig. 7. Packet error ratio (LVR=0).

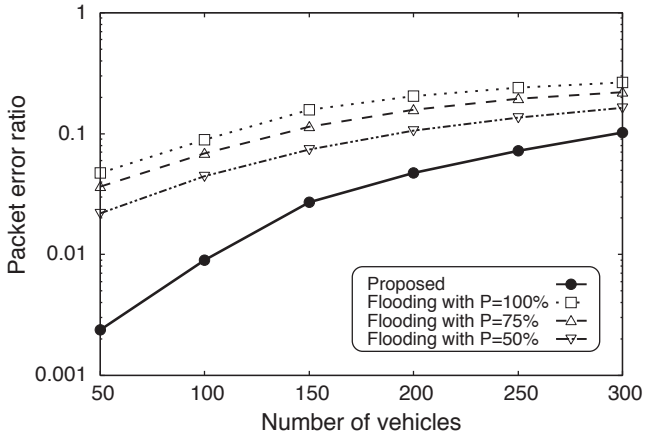


Fig. 8. Packet error ratio (LVR=0.2).

the results, we can find that our proposed scheme can achieve the highest delivery ratio. The performance of the full flooding scheme degrades when the number of vehicles increases. This reason is that broadcast storm problems tend to occur when the number of vehicles increases. Meanwhile, the delivery ratio of the probabilistic flooding scheme degrades when the flooding probability decreases. Especially, it degrades much when the value of the flooding probability is set to low. This is because several vehicles are required to forward vehicle information messages when there are a small number of vehicles on the road. Additionally, the performance of the every mechanism degrades when the number of vehicles is set to 50 or 100. The reason is that there are not enough vehicles to forward vehicle information messages in the delivery area.

Figure 6 shows the area delivery ratio of vehicle information messages with the large-size vehicle ratio equals to 20 [%]. Therefore, blocking due to large-size vehicles is considered in this condition. From the results, the proposed scheme can keep the high delivery ratio, which is almost same as that of Fig. 5. On the contrary, the performance of the flooding scheme degrades compared with Fig. 5. This reason is that the proposed scheme can obtain the multi-path diversity effect. Therefore, our scheme has resistance to blocking.

Figure 7 shows the packet error ratio of vehicle information messages with the large-size vehicle ratio equals to 0 [%]. From the results, the proposed scheme can reduce the packet error ratio compared with the flooding schemes. The reason is that the proposed scheme can demodulate some same OFDM signals. Therefore, vehicles tend to obtain better condition OFDM signals even if condition of OFDM signals degrades due to fading or blocking. Meanwhile, the packet error ratio increases with increasing in the number of vehicles. This is because interference signals also increase according to increasing in the number of transmitted packets.

Figure 8 shows the packet error ratio of vehicle information messages with the large-size vehicle ratio equals to 20 [%]. From the results, we can find that the packet error ratio is improved compared with Fig. 7. This is caused by large-size

 ISBN:978-1-61208-113-7

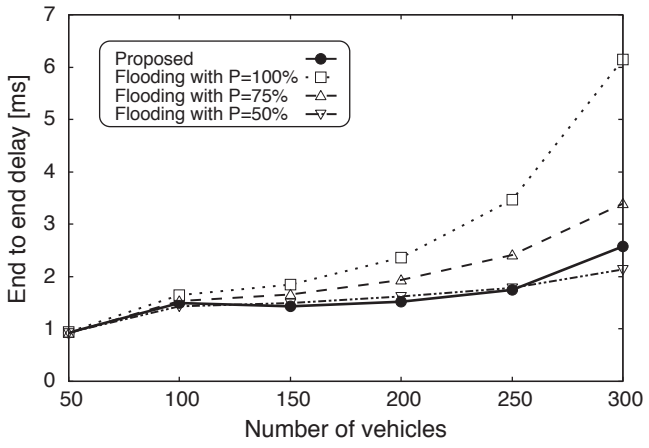


Fig. 9. End-to-end delay (LVR=0).

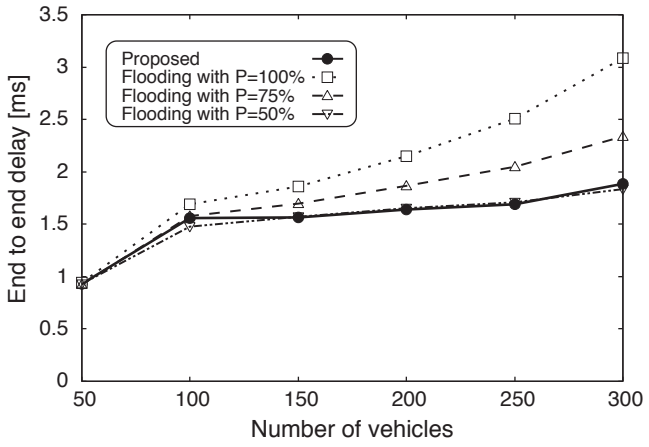


Fig. 10. End-to-end delay (LVR=0.2).

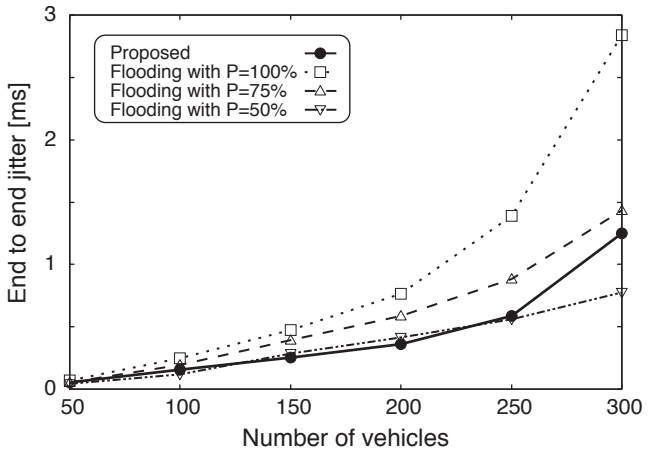


Fig. 11. End-to-end jitter (LVR=0).

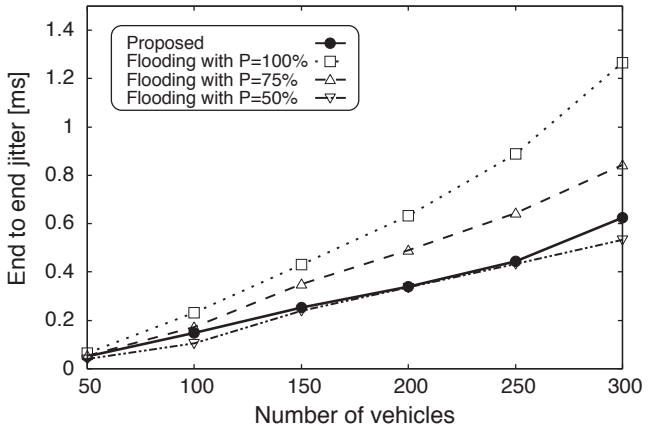


Fig. 12. End-to-end jitter (LVR=0.2).

vehicle can also block interference signals. Hence, SINR is improved due to reduction of interference power.

Figure 9 shows the end-to-end delay performance with the large-size vehicle ratio equals to 0 [%]. The delay period starts when a source vehicle transmits a vehicle information message, and ends when the vehicle information message is received at a vehicle in the delivery area. Therefore, the accurate delay of each vehicle is different due to the positions of the vehicles. Therefore, the delay performance averages delays of each vehicle in the delivery area. From the results, we can find that the proposed scheme can reduce the delay to the halves of the full flooding scheme. The reason is that vehicles transmit the same OFDM signal simultaneously in the proposed scheme. Hence, vehicles tend to obtain packet transmission opportunity according to reduction of consumed wireless resource. In road safety applications, delay is the most important factor to achieve collision avoidance and control of traffic flows. Moreover, our scheme can reduce the increasing amount of delay according to the increasing in the number of vehicles. Hence, our scheme has scalability for number of vehicles.

Figure 10 shows the end-to-end delay performance with the large-size vehicle ratio equals to 20 [%]. The results show that the proposed scheme can reduce the delay performance compared with Fig. 9. The reason is that blocking by large-size vehicles improves frequency reuse performance. Then, vehicles tend to obtain packet transmission opportunities due to reduction of reception signals.

Figure 11 shows the end-to-end jitter performance with the large-size vehicle ratio equals to 0 [%]. From the results, we can find that the proposed scheme can also reduce the jitter to the halves of the full flooding scheme. Moreover, the proposed scheme can keep low jitter performance even if the high area delivery ratio is maintained.

Figure 12 shows the end-to-end jitter performance with the large-size vehicle ratio equals to 20 [%]. The results show that the proposed scheme can reduce the jitter performance due to the same reason in the delay performance.

 ISBN:978-1-61208-113-7

V. Conclusion

In this paper, we have focus on characteristics of OFDM communication that vehicles can demodulate some same OFDM signals in guard interval period. The proposed scheme offers the autonomous media access control scheme that some vehicles can transmit the same signal at same instance. Moreover, we have employed the proposed media access control scheme and the broadcast mechanism to achieve data dissemination of vehicle information messages. From the results, our scheme can keep the high message delivery ratio and the low end-to-end delay even if fast movement of vehicles, blocking by large-size vehicles and channel fluctuation due to fading are considered. In the actual communication environment, it is important to support these mixed factors for real safe driving systems. Meanwhile, we can provide required quality in communication if we employ the forward error correction (FEC) to recover the packet losses. Considering all these results mentioned above, the proposed scheme could be one of new fundamental schemes for achieving ITS.

Acknowledgment

This work is supported in part by Telecommunications Advancement Foundation of Japan.

References

[1] V. Naumov, R. Baumann, T. Gross, "An evaluation of inter-vehicle ad hoc networks based on realistic vehicular traces," ACM MobiHoc '06, pp. 108–119, May 2006.

[2] J. Mittag, F. Thomas, J. Härri, and H. Hartenstein, "A comparison of single- and multi-hop beaconing in VANETs," ACM VANET '09, pp. 69–78, Sep. 2009.

[3] S. Y. Wang, "On the effectiveness of distributing information among vehicles using inter-vehicle communication, " IEEE Intelligent Transportation Systems 2003, Vol. 2, No. 12–15, pp. 1521–1526, Oct. 2003.

[4] F. Gil-Castineira, F.J. Gonzalez-Castano, and L. Franck, "Extending Vehicular CAN Fieldbuses With Delay-Tolerant Networks," IEEE Transactions on Industrial Electronics, Vol. 55, No. 9, pp. 3307–3314, Sep. 2008.

[5] Y. Toor, P. Muhlethaler, A. Laouiti, and A. de La Fortelle, "Vehicle Ad Hoc networks: applications and related technical issues," IEEE Communications Surveys and Tutorials, Quarter 2008, Vol. 10, No 3, pp. 74–88, 2008.

[6] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing," IETF Request for Comments 3561, Jul. 2003.

[7] D. Johnson, Y. Hu, and D. Maltz, "The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4," IETF Request for Comments 4728, Feb. 2007.

[8] F. Li and Y. Wang, "Routing in vehicular ad hoc networks: A survey," IEEE Vehicular Technology Magazine, Vol. 2, No. 2, pp. 12–22, Jun. 2007.

[9] R.A. Santos, A. Edwards, R.M. Edwards, and N.L. Seed, "Performance evaluation of routing protocols in vehicular ad-hoc networks, " International Journal of Ad Hoc and Ubiquitous Computing, Vol. 1, No. 1-2, pp. 80–91, 2005.

[10] H. Hartenstein, B. Bochow, A. Ebner, M. Lott, M. Radimirsch, and D. Vollmer, "Position-aware ad hoc wireless networks for inter-vehicle communications: the Fleetnet project, " ACM international symposium on Mobile ad hoc networking & computing (MOBIHOC 2001), pp. 259–262, Oct. 2001.

[11] Z. Mo, H. Zhu, K. Makki, and N. Pissinou, "MURU: A Multi-Hop Routing Protocol for Urban Vehicular Ad Hoc Networks, " International Conference on Mobile and Ubiquitous Systems: Networks and Services (MOBIQUITOUS 2006), Jul. 2006.

[12] F. Granelli, G. Boato, and D. Kliazovich, "MORA: a Movement-Based Routing Algorithm for Vehicle Ad Hoc Networks, " IEEE Workshop on Automotive Networking and Applications (AutoNet 2006), Dec. 2006.

[13] V. Naumov and T.R. Gross, "Connectivity-Aware Routing (CAR) in Vehicular Ad-hoc Networks, " IEEE International Conference on Computer Communications (INFOCOM 2007), pp. 1919–1927, May 2007.

[14] Q. Yang, A. Lim, S. Li, J. Fang and P. Agrawal, "ACAR: Adaptive Connectivity Aware Routing for Vehicular Ad Hoc Networks in City Scenarios," MOBILE NETWORKS AND APPLICATIONS, Vol. 15, No. 1, pp. 36–60, Feb. 2010.

[15] T.D.C. Little and A. Agarwal, "An information propagation scheme for VANETs," IEEE Intelligent Transportation Systems (ITSC 2005), pp. 155–160, Sep. 2005.

[16] Y. Mylonas, M. Lestas, A. Pitsillides, "Speed adaptive probabilistic flooding in cooperative emergency warning," ACM WICON '08, No. 81, Nov. 2008.

[17] Y. Chen, Y. Lin, S. Lee, "A Mobicast Routing Protocol in Vehicular Ad-Hoc Networks," IEEE Globecom 2009, pp. 1–6, Nov. 2009.

[18] Y.-C. Tseng, S.-Y. Ni, Y.-S. Chen, and J.-P. Sheu, "The broadcast storm problem in a mobile ad hoc network, " Wireless Networks, Vol. 8, pp. 153–167, 2002.

[19] W. Lou and J. Wu, "On reducing broadcast redundancy in ad hoc wireless networks, " IEEE Transactions on Mobile Computing, Vol. 1, No. 2, pp. 111–122, 2002.

[20] Y.-C. Tseng, S.-Y. Ni, E.-Y. Shih, "Adaptive approaches to relieving broadcast storms in a wireless multihop mobile ad hoc network, " IEEE Transactions on Computers, Vol. 52, No. 5, pp. 545–557, May 2003.

[21] E. Fasolo, A. Zanella, and M. Zorzi, "An Effective Broadcast Scheme for Alert Message Propagation in Vehicular Ad hoc Networks, " IEEE International Conference on Communications (ICC '06), pp. 3960–3965, Jun. 2006.

[22] S. Eichler, "Performance Evaluation of the IEEE 802.11p WAVE Communication Standard," IEEE VTC 2007-Fall, pp. 2199–2203, Sep. 2007.

[23] K. Bilstrup, E. Uhlemann, E. G. Strom, and U. Bilstrup, "Evaluation of the IEEE 802.11p MAC Method for Vehicle-to-Vehicle Communication," IEEE VTC 2008-Fall, pp. 1–5, Sep. 2008.

[24] D. Lee and K. Cheun, "A new symbol timing recovery algorithm for OFDM systems," IEEE Transactions on Consumer Electronics, Vol. 43 No. 3, pp. 767–775, Aug. 1997.

[25] L. Thibault and M. T. Le, "Performance evaluation of COFDM for digital audio broadcasting. I. Parametric study," IEEE Transactions on Broadcasting, Vol. 43, No. 1, pp. 64–75, Mar. 1997.

[26] QualNet, URL:http://www.scalable-networks.com

 ISBN:978-1-61208-113-7

An Adaptive Mechanism for Access Control in VANETs

Alisson Barbosa de Souza, Ana Luiza Bessa de P. Barros, Antônio Sérgio de S. Vieira, Filipe Maciel Roberto, Joaquim Celestino Júnior
Computer Networks and Security Laboratory (LARCES)
State University of Ceará (UECE)
{alisson, analuiza, sergiosvieira, filipe, celestino}@larces.uece.br

***Abstract*—VANETs have the ability to transmit various types of information between a vehicle and another, or between a vehicle and a fixed station. However, transmission may be impaired due to long delays, quantity of collisions and signal noise. This is caused, in part, by abrupt changes in the network topology. Thus, in order to maintain the quality and stability of the network, a mechanism is proposed for the vehicle to self-adapt dynamically according to the context. For this purpose, MAC layer parameters must be changed, allowing better control of access to the medium.**

***Keywords*—VANET; 802.11p; Contention Window; Density; Fuzzy Logic.**

I. INTRODUCTION

The main goal of VANETs (Vehicular Ad Hoc Networks) is to allow vehicles to exchange information to enable the use of safety support applications such as: emergency warning systems and accident prevention. But, comfort applications are another type of application that will be present in VANETs [1].

The network environment present in VANETs can take on diverse configurations. It can be a sparse network without atmospheric interference, or it can be a dense network on a rainy day in a big city. This second scenario can lead to a strong degradation of network performance. Moreover, this environment can change from sparse to dense or vice-versa. In each case, the network needs to adapt to the environment in order to work properly [2].

Thus, in order to optimize network resources, meet the requirements of different applications and dynamically adapt to network conditions and traffic, we propose a mechanism that has been developed with the support of fuzzy intelligence in order to better control the VANETs and adapt medium access control (MAC).

This work is organized as follows: Section 2 presents related works. In Section 3 the theoretical basis is shown. The architecture is explained in detail in Section 4. Sections 5 and 6 report on the scenario, results and analysis. Future works and conclusion are outlined in the last sections.

II. RELATED WORK

In VANETs, the adaptability of protocols to the environment has been investigated with the goal of not letting a highly changeable scenario degrade network quality.

Shankar *et al.* [3] shows that the rapid changes in the quality of connections and the rapid mobility of vehicles cause the sub-utilization of network resources when the default network settings are static. Thus, a scheme is proposed for adapting the transmission rate to better utilize network capacity. For this adaptation, the proposed scheme evaluates some information from GPS (Global Positioning System) and some metrics of network performance. However, in this paper, density was not used as a context parameter.

Artimy *et al.* [4], consider density as an important parameter in their work. The rapid change in topology, due to traffic jams, is shown to disturb the homogenous distribution of vehicles on the road. Dynamic transmission power has been proposed as a manner to maintain network connectivity and minimize the adverse effects of unregulated power.

The contention window (CW) also plays an important role in adjusting the network. In the paper of Wang *et al.* [5], the contention window (CW) of a vehicle is adapted according to the neighborhood density of a stationary unit (RSU). Thus, the adaptation algorithm is centered on the RSU.

Another aspect that has received attention in the scientific community is Quality of Service (QoS), because it is very vulnerable to the environment. Adler *et al.* [6] proposes a system to prioritize messages based on context and content. On this basis, a function of relevance is calculated for each message, and each message will have different CWs (Contention Window).

Protocol scalability also gains with the adaptation. Mertens *et al.* [7] states that, in broadcast scenarios, there are significant problems with scalability due to the flood of messages among the cars. In this adaptation, the CW is modified according to the PER (Packet Error Rate), and the data transmission rate is changed according to the degree of congestion of the channels.

It is important to quote the all the approaches proposed above for VANETs MAC layer address changes have a reactive aspect, i.e., they wait for the channel to become congested or for the level of the PER to increase. These solutions do not prevent large amounts of packets from being discarded while waiting for the network to adapt through some mechanism in order to change that scenario. Another unpromising approach is to individually set the importance of each message through different CWs for each one. This creates a large overhead and processing time.

In order to overcome this problem, this article proposes a mechanism of network adaptation to the traffic scenario in

 ISBN:978-1-61208-113-7

VANETs in order to improve the quality of transmission and reception of packets. Traffic density information are taken into account as a parameter to be used by a traffic based fuzzy logic analyzer and, thereafter, the MAC layer parameters can be change dynamically. The proposed architecture consists of two main modules: a contextual information captor and an information analyzer. It is important to note that the changes are predictive, since they do not have to wait for any degradation of the network to make any adjustments.

III. Theoretical Foundation

A. Density

In different VANETs scenarios, speed fluctuation, traffic signs, the road model and other factors that are described in traffic engineering contribute significantly to changes in network density, disrupting homogeneous node distribution. These abrupt and frequent changes create a highly dynamic topology and can cause degradation in network performance if the protocols are not designed to handle such situations.

Panichpapiboon *et al.* [8] and Yousefi *et al.* [9] show studies on network connectivity and attest to the importance of density for the connection. Among other things, they emphasize that density affects network connectivity proportionally, i.e., the higher the density, the greater the connectivity. The impact of network connectivity can be felt in different ways. Tonguz *et al.* [10] shows that higher the density, the higher the packet loss rate due to problems of contention and collisions. Moreover, reducing the density increases idle time spent in transmission / reception of a packet and for a sparse network there will be more retransmissions as observed in Shankar *et al.* [3] and was observed that increasing the density increases the rate of penetration, i.e., the number of nodes that can be reached by the message [8] [10].

Therefore, density can be considered a very important feature for VANETs. A protocol project should consider its influence on the quality of transmissions to allow continuous and reliable exchange of information between vehicles.

B. Backoff Time

Backoff time is a time value that determines the time of transmission. It is calculated by a random value, chosen based on the contention window, multiplied by a time slot [11]. Higher priority will be assigned to the least amount of backoff time. The backoff value has been taken into account in the fuzzy system calculations and in the evaluations made in this work.

Besides the density, Natkaniec *et al.* [12] estimated that the contention window is crucial to reduce the probability of collision and increase network throughput. Bianchi [13] presents a more detailed study of this feature by showing that the saturation throughput, limit reached by the throughput in overcrowded conditions, is strongly dependent on the contention window and that its optimal value choice depends on the number of network nodes.

C. Fuzzy Logic

The probability theory can be used to formally represent information in stochastic decision environments. It represents the uncertainty associated with the randomness of events. The theory of fuzzy sets, in turn, seeks to represent the uncertainty associated with vague, inaccurate or independently unrelated information. These sets were developed by Lotfi Zadeh and initially published in 1965 [14].

Given that present day complex networks are dynamic, i.e., there is great uncertainty associated with the input traffic and other environmental parameters, that they are subject to unexpected overloads, failures and disturbances, and that they defy accurate analytical modeling, fuzzy logic appears to be a promising approach to address key aspects of networks. The ability to model networks in the continuum mathematics of fuzzy sets rather than with traditional discrete values, coupled with extensive simulation, offers a reasonable compromise between rigorous analytical modeling and purely qualitative simulation [15].

D. Dedicated Short Range Communications (DSRC)

Dedicated Short Range Communications (DSRC) is the standardization of a spectrum band in the United States [2]. In 1999, the Federal Communication Commission of the United States allocated 75MHz of the DSRC spectrum in 5.9 GHz to be used exclusively for vehicle-to-vehicle or vehicle-to-infrastructure communications.

The DSRC spectrum is divided into 7 channels of 10 MHz. Channel 178 is the control channel (CCH), which is exclusive for security communications. The two side channels are reserved for special uses. Others are service channels (SCH) available for use in security and comfort communications.

WAVE architecture (Wireless Access in the Vehicular Environment) uses standard DSRC.

E. Wireless Access in the Vehicular Environment

In 2004, the IEEE began the standardization of communications in vehicular networks, called WAVE architecture (figure 1) which is defined currently in five documents: IEEE P1609.1, IEEE P1609.2, IEEE P1609.3, IEEE P1609.4, IEEE 802.11p.

IEEE 802.11p defines the physical layer and medium access control (MAC) for vehicular networks. This proposed standard specifies the extensions to IEEE 802.11 that are necessary to provide wireless communications in a vehicular environment.

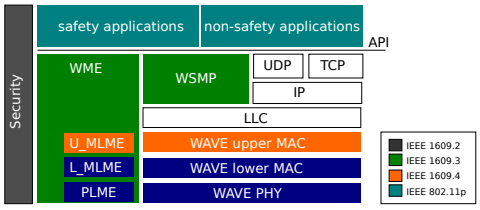


Fig. 1. WAVE Architecture

F. Wave Short Message Protocol (WSMP)

According to Figure 1, WSMP (WAVE Short Message Protocol) is an alternative to the use of TCP / UDP, and IPv6 in WAVE environments. The justification of an alternative network service is the greater efficiency in the WAVE environment, where it is expected that most applications require very low latency and are non-connection-oriented. Many broadcast applications use WSMP to minimize the size of messages and reduce the delay for critical security messages [16] [17].

IV. Proposed Mechanism

In this paper, we propose a mechanism for backoff time self-adaptation in VANETs, as illustrated in Figure 2. The Captor module is responsible for obtaining density information that is passed to the Analyzer. The Analyzer also receives information about the value of the random backoff time chosen. from the 802.11p MAC layer protocol. Using the value received, the analyzer will check if this value is conforming to the current situation of vehicular traffic. Without this mechanism, in a very dense scenario, the value of the backoff time chosen may be very small. In this case, when there are several transmitting cars, there may be a higher probability of collisions, losses, etc.

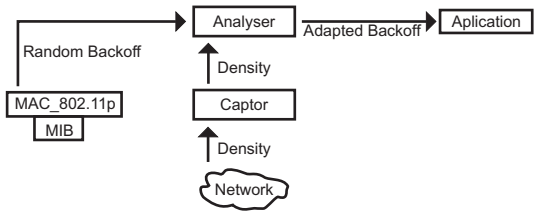


Fig. 2. Proposed Mechanism

One advantage of this mechanism is the use of prediction rather than reaction. There is no need to wait for the degradation of the service network to do something about it. In reactive systems, during the time taken to measure the degree of congestion in a channel, the amount of PER or the amount of collisions, only to adapt the network parameters, there may be more collisions and more packet loss. Using density as a context descriptor and performing dynamic update of parameters, there is no need to let the network conditions get worse and then make an adjustment. Before there is a decline in transmission quality, the network can already adapt and thus maintain its stability.

A. Captor

Literature mentions two ways to obtain the density for a particular vehicle. The density can be disseminated among the vehicles through beacon messages [3] [18]. Each vehicle delivers its speed and position to other vehicles. Thus, a vehicle can count how many neighbors are in its range and calculate its own density, and it may also spread its own density. Another way is mentioned in Artimy *et al.* [4]. The density is estimated based on the number and length of stops the vehicle makes. The more the car stops, and the longer it stands there, the greater the density.

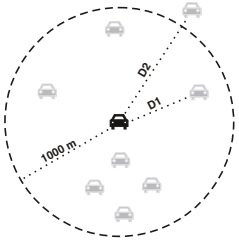


Fig. 3. Density by Transmission Range

As the focus of this paper is not to find the best way to get the density of a vehicle for this task we used a native function of the simulator (see Section 5). The function calculates the distance between vehicles in a given transmission range. As the distance from a node to a given vehicle diminishes in relationship to the transmission range of that vehicle, the value of the density is increased. In Figure 3, the central vehicle has a density of 7. The vehicle is at a distance D1 is being recorded for that value, but the car at a distance D2 is not, because D2 is greater than the central vehicle transmission range.

The context information need not be calculated for each transmission [3]. Considering that in scenarios of high mobility (vehicular speeds equal to 105km / h) in 100ms a vehicle has moved less than 3m. As the density changes little in that interval, we used 1 second periods to capture information about density.

B. Analyzer

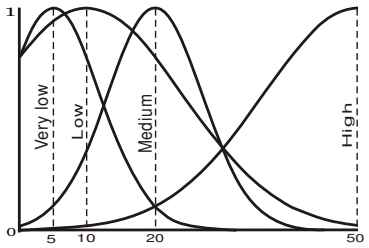


Fig. 4. Fuzzy Sets for Density

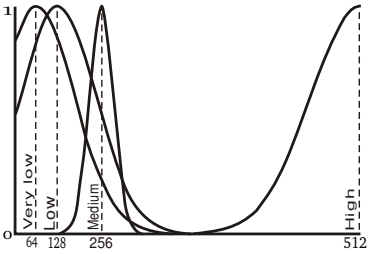


Fig. 5. Fuzzy Sets for Backoff Value

The Analyzer receives the following information: the density, from the Captor module, and the backoff time based on

the MIB contention window range defined by the 802.11p protocol. The purpose of this module is to ensure that the backoff time will be adjusted in accordance with the traffic scenario.

A fuzzy system is used to describe the values. Thus, a very dense network should increase its backoff time to attempt a reduction in the amount of packet collisions. A sparse network should reduce the backoff time in order not to underutilize network resources.

Figure 4 shows the fuzzy sets for density classification based on Wisitpongphan *et al.* [19].

Figure 5 shows the fuzzy sets that return the optimal backoff time based on Wang *et al.* [5].

The fuzzy system calculates a new backoff value based on vehicle density in order to optimize access to the medium. If the density is low, the fuzzy system returns a small backoff value and close to the optimum value [5] in order not to underutilize the network. On the other hand, if the network is dense, the returned backoff value is great and near to optimal value, enough to provide a good cost-benefit relationship between throughput and collisions or loss.

A new backoff value is generated only when the channel is busy, in other words, when the node is not transmitting, although the density values are sent every second to the Analyzer.

However, to avoid all vehicles from calculating the same backoff time, the adjusted value is added to the little random value generated by the 802.11p MAC layer protocol, which is passed to the analyzer.

V. Scenario

The simulator used for the experiments was the NCTUns 6.0 [20]. The NCTUns has a complete implementation of IEEE 1609 and 802.11p standards. It is opensource and is allowed to add on new modules and agents. However, the simulator does not support more than 4096 nodes in a simple simulation.

The scenario used was a stretch of highway 6km long. All vehicles have OBUs (On Board Units) and there is no RSU (Road Side Unit). The propagation model used was the Two Ray Ground [21]. Vehicular traffic was generated according to a Poisson process. Three environments were tested in this scenario: a sparse density environment with 35 vehicles, an average density environment with 50 vehicles, and a highly dense environment with 200 vehicles.

Each vehicle has a transmission range of 1km and its mobility is controlled automatically. This control is carried out by an agent attached to the simulator called *CarAgent*. The vehicles have a maximum speed of 130km/h, maximum acceleration of $3m/s^2$ and maximum deceleration of $5m/s^2$. All vehicles have 1.5 meter omnidirectional antennas.

Traffic is generated through an agent called WSM which simulates WSMP, but without retransmissions. Every 100ms, a broadcast Wave Short Message is transmitted. For experiments, we used an application that works with WSMP. Each WSMP message has a length of 1458 bytes and uses the control channel (178).

VI. Results and Analysis

The aim of the experiments is to compare the adaptive approach with the non-adaptive approach, i.e., to compare the dynamic approach with the 802.11p standard. According to Bilstrup *et al.* [11], IEEE 802.11p uses the following to calculate backoff: (i) it chooses a uniform distribution integer between 0 and $CWmin$ (minimum contention window for a given class), (ii) it multiplies the choosen integer by a certain time slot at the physical layer, (iii) it decreases the backoff only when the channel is free, (iv) when backoff reaches zero, it transmits immediately. When a problem is detected in the transmission, the value of the contention window is doubled. Upon successful transmission, the contention window value returns to the initial value. However, in *broadcast* situations, there is no way to know if there was problem in the transmission because there is no confirmation. Thus, the contention window value is always CWmin. Thus, there is a higher probability of calculating the same backoff time and transmitting data at the same time, causing a greater number of collisions [22].

To analyze the network situation, we used three metrics: number of packets received per second (BRX), amount of packet loss (DROP) and percentage of success (SUC). The amount of packets lost is the sum of errors caused by collisions and discards. The success rate is the number of packets received divided by the number of packets that should have been successfully received if no losses occurred. Thus, $SUC = BRX/(BRX + DROP)$.

In the drop chart and packet loss chart, these values grow up to a certain point, and then they decrease. This happens because the densities calculated for each vehicle are very low in the first and the last moments. The cars enter and exit by the Poisson process. These metrics will be highest when all vehicles are present on the track, when it has the highest density.

In figures 6, 7 and 8, with the adaptive approach, packet loss is less than with the standard approach. This happens because the vehicles use a backoff value close to optimum, reducing the likelihood of the medium being used at the same time.

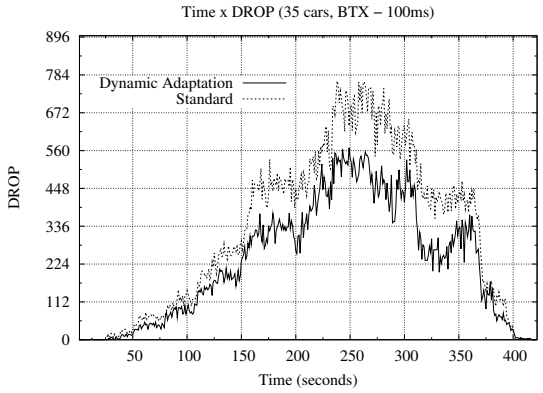


Fig. 6. Quantity of overall packet losses for sparse scenarios

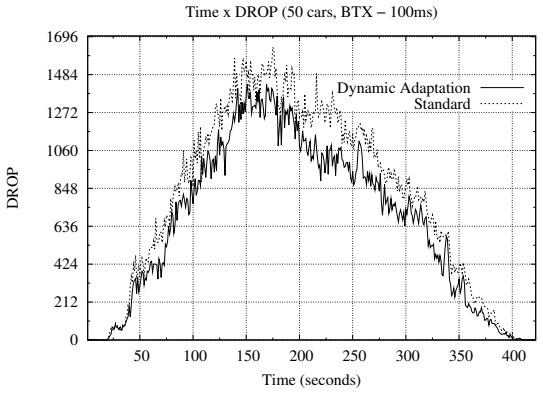


Fig. 7. Quantity of overall packet losses for medium scenarios

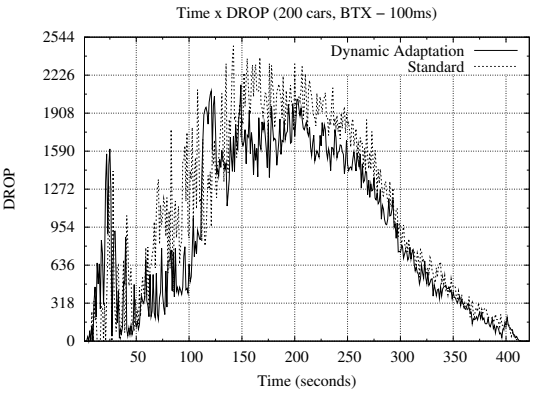


Fig. 8. Quantity of overall packet losses for dense scenarios

In figures 9, 10 and 11 with the adaptive approach, receptions per second is greater than with the standard approach, i.e., a greater number of neighboring vehicles is reported from a single message from a transmitting node. Since there is no retransmission, this metric provides a view of the rate of packet delivery in the neighborhood of a node. Moreover, backoff time is close to the optimum value. This way, the medium is better shared.

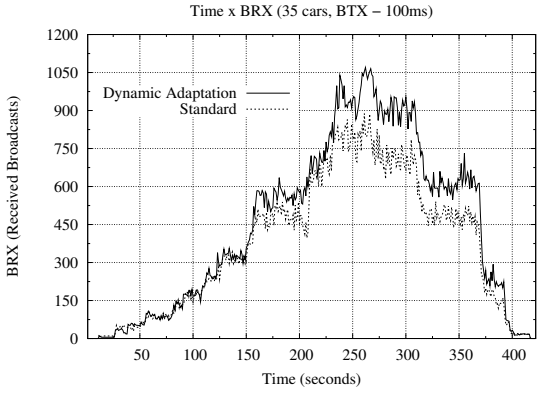


Fig. 9. Overall quantity of received packets per second for sparse scenarios.

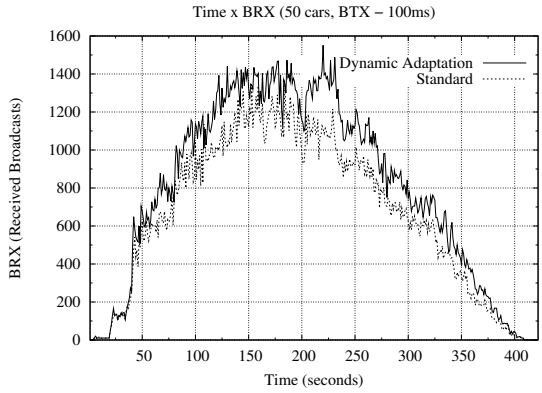


Fig. 10. Overall quantity of received packets per second for medium scenarios.

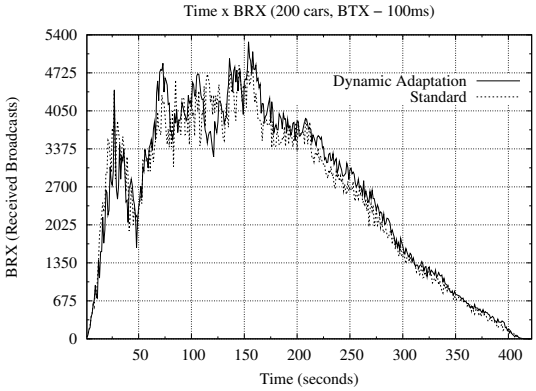


Fig. 11. Overall quantity of received packets per second for dense scenarios.

In figures 12, 13 and 14 with the adaptive approach, there is an improvement in the percentage of success because this approach received more packets per second with less losses. This means that, considering the total number of packets that should be received, the adaptive approach was more successful than the standard approach.

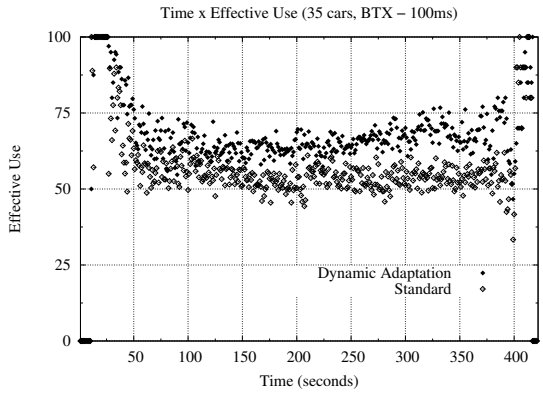


Fig. 12. Percentage of success for sparse scenarios

Therefore, the adaptive approach provides better network quality by adjusting backoff time and optimizing sharing of

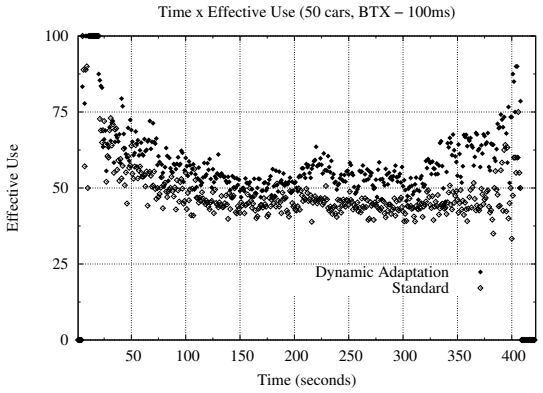


Fig. 13. Percentage of success for medium scenarios

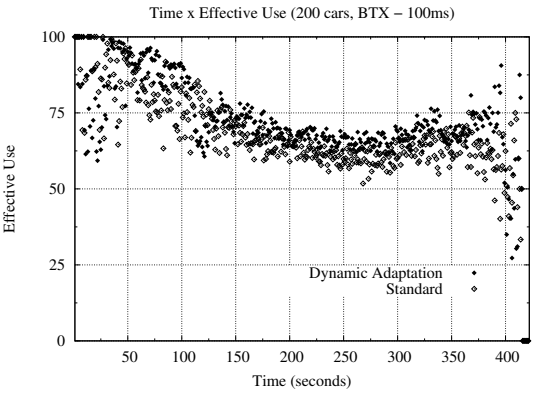


Fig. 14. Percentage of success for dense scenarios

the wireless medium. Thus, with the adaptive approach, we obtained better performance than with the standard approach (802.11p).

VII. Conclusion and Future Work

This work, proposed a mechanism for context adaptation to better control network use. Density was used as a context descriptor, and the backoff time as a parameter to be changed dynamically in each vehicle to handle its access to medium.

The adaptive approach has proven effective for all scenario types: sparse, medium and dense. Collisions and drops decreased. Useful network throughput was increased since the amount of packets received per second increased in all scenarios evaluated.

In the future, we intend to work with other network parameters, such as data rate, transmission power, AIFS, etc. Other context parameters as speed, acceleration, connectivity, signal strength, BER, etc can also be verified. In addition, applications that use TCP, UDP or RTP will be tested. Other ways to obtain density will also be analyzed.

References

[1] S. Yousefi, M. Mousavi, and M. Fathy, "Vehicular ad hoc networks (VANETs): challenges and perspectives," in *ITS Telecommunications Proceedings, 2006 6th International Conference on*, 2006, pp. 761–766.

[2] D. Jiang and L. Delgrossi, "IEEE 802.11 p: Towards an international standard for wireless access in vehicular environments," in *IEEE Vehicular Technology Conference, 2008. VTC Spring 2008*, 2008, pp. 2036–2040.

[3] P. Shankar, T. Nadeem, J. Rosca, and L. Iftode, "CARS: Context-Aware Rate Selection for Vehicular Networks," in *The sixteenth IEEE International Conference on Network Protocols (ICNP 2008)*, 2008, pp. 19–22.

[4] M. Artimy, W. Robertson, and W. Phillips, "Assignment of dynamic transmission range based on estimation of vehicle density," in *Proceedings of the 2nd ACM international workshop on Vehicular ad hoc networks*. ACM New York, NY, USA, 2005, pp. 40–48.

[5] Y. Wang, A. Ahmed, B. Krishnamachari, and K. Psounis, "IEEE 802.11 p Performance Evaluation and Protocol Enhancement," in *IEEE International Conference on Vehicular Electronics and Safety, 2008. ICVES 2008*, 2008, pp. 317–322.

[6] C. Adler, R. Eigner, C. Schroth, and M. Strassberger, "Context-Adaptive Information Dissemination in VANETs Maximizing the Global Benefit," in *Fifth IASTED International Conference on Communication Systems and Networks (CSN 2006)*, 2006.

[7] Y. Mertens, M. Wellens, and P. Mahonen, "Simulation-Based Performance Evaluation of Enhanced Broadcast Schemes for IEEE 802.11-Based Vehicular Networks," in *IEEE Vehicular Technology Conference, 2008. VTC Spring 2008*, 2008, pp. 3042–3046.

[8] S. Panichpapiboon and W. Pattara-atikom, "Connectivity Requirements for Self-Organizing Traffic Information Systems," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3333–3340, 2008.

[9] S. Yousefi, E. Altman, and R. El-Azouzi, "Study of connectivity in vehicular ad hoc networks," in *Workshop on Spatial Stochastic Models in Wireless Networks (SpasWin2007), Cyprus*, 2007, pp. 573–587.

[10] O. Tonguz, N. Wisitpongphan, F. Bai, P. Mudalige, and V. Sadekar, "Broadcasting in VANET," *2007 Mobile Networking for Vehicular Environments*, pp. 7–12, 2007.

[11] K. Bilstrup, E. Uhlemann, E. Strom, U. Bilstrup, and O. Altintas, "On the ability of the 802.11 p MAC method and STDMA to support real-time vehicle-to-vehicle communication," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, 2009.

[12] M. Natkaniec and A. Pach, "An analysis of the backoff mechanism used in IEEE 802.11 networks," in *Proc. of fifth IEEE symposium on Computers and Communications*, 2002, pp. 3–6.

[13] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on selected areas in communications*, vol. 18, no. 3, pp. 535–547, 2000.

[14] L. Zadeh, "Fuzzy sets," *Fuzzy Systems and AI Reports and Letters*, vol. 61, pp. 129–136, 1965.

[15] Q. R. H. J. S. S. Ghosh and A. Celmins, "A survey of recent advances in fuzzy logic in telecommunications networks and new challenges," *IEEE Transactions on Fuzzy Systems*, vol. 6(3):443–447, 1998.

[16] IEEE, "IEEE P1609.3/D22 Trial-use Standard for Wireless Access in Vehicular Environments (WAVE) - Networking Services," *Draft IEEE Standard*, 2007.

[17] R. Alves, I. Campbell, R. Couto, M. Campista, I. Moraes, M. Rubinstein, L. Costa, O. Duarte, and M. Abdalla, "Redes Veiculares: Princípios, Aplicações e Desafios," *Minicursos do Simpósio Brasileiro de Redes de Computadores, SBRC*, 2009.

[18] B. Bako, I. Rikanovic, F. Kargl, and E. Schoch, "Adaptive Topology Based Gossiping in VANETs Using Position Information," *Lecture Notes in Computer Science*, vol. 4864, p. 66, 2007.

[19] N. Wisitpongphan, O. Tonguz, J. Parikh, P. Mudalige, F. Bai, V. Sadekar *et al.*, "Broadcast storm mitigation techniques in vehicular ad hoc networks," *IEEE Wireless Communications*, vol. 14, no. 6, p. 84, 2007.

[20] (2009) Website of NCTUns 6.0 Network Simulator and Emulator. [Online]. Available: http://nsl.csie.nctu.edu.tw/nctuns.html

[21] N. Eude, B. Ducourthial, and M. Shawky, "Enhancing ns-2 simulator for high mobility ad hoc networks in car-to-car communication context," in *Proceedings of the 7th IFIP International Conference on Mobile and Wireless Communications Networks (MWCN 2005), Morocco*, 2005.

[22] M. Koubek, S. Rea, and D. Pesch, "Effective Emergency Messaging in WAVE based VANETs," *Proceedings of the 1st International Conference on Wireless Access in Vehicular Environments (WAVE '08), Dearborn, USA*, 2008.

 ISBN:978-1-61208-113-7

Methodology of Dynamic Architectural Adaptation for Ad hoc Networks Operating in Disturbed Environment

Farouk Aissanou
Université du Québec en Outaouais
Gatineau, Canada
aisf01@uqo.ca
Dept. RS2M, TELECOM Sud
Paris, France
Farouk.aissanou@it-sudparis.eu

Ilham Benyahia
Dept. Of Computing and Engineering
Université du Québec en Outaouais
Gatineau, Canada
benyahia@uqo.ca

Abstract— **Wireless networks and particularly ad hoc networks, are gaining in speed and capacity. These advances open the way to their use in emergent, increasingly complex applications. Such networks have to operate in disturbed environments where disturbances, mainly caused by fading and interferences, primarily originate by the physical layer. Congestions associated with environment-specific disturbances caused by complex applications, such as emergency and disaster applications, are the second source of disturbance. Such networks must guarantee a QoS (Quality of Service) management to their associated applications, a task that is possible only by minimizing the transmission delay and maximizing the packets delivery ratio. The conventional network architecture used for TCP/IP model shows degradations of performance, especially when the networks operate in environments with physical layer disturbances. This paper presents a study based on the routing performance of ad hoc networks operating in disturbed environment. Simulation results are presented and analyzed to illustrate limitations of the conventional ad hoc network architecture. The methodology of a network architectural design based on cross-layer architecture using a multi-criteria decision making process for quality enhancement is also presented. This work enabled us to highlight a new direction for the communication architecture of cognitive vehicular networks operating under disturbed environment. This direction consists of considering a dynamic reconfiguration of the communication architecture according to the network environment behaviour. Thus, it will allow either the traditional architecture or the cross layer architecture based on autonomous components.**

Keywords- ad hoc networks; ant colony optimization algorithm; adaptive network communication architecture; cross-layer architecture; multi-criteria decision making.

I. INTRODUCTION

Wireless networks, including ad hoc networks, are among emergent telecommunication technologies operating in environments that can cause link failures and degradation of QoS (Quality of Service) in their associated applications. Despite the advances of such technologies, such as large band for G3 and G4 (third and fourth generations of mobile standards and technology), there is still a risk of performance degradation. This risk is especially significant for constrained services such as those used for real-time or near real-time applications. The correlation between physical layer phenomena of ad hoc networks and constraints related to complex applications that involve close nodes and rapid changes in topology make the behaviour of ad hoc network very complex and hard to predict [1][2][3][4]. The effect of this correlation is the guarantee of only the minimum required QoS (Quality of Service). This paper presents an analysis of conventional network architectures operating in disturbed environment. It offers a model for the cross-layer concept that makes it possible to: 1) enhance the network performance by sharing information between non adjacent layers and, 2) take benefits of real-time notifications of channels events representing new states that may have an impact on the performance of network protocols, which also include routing protocols.

Section 2 presents the application context of emergent telecommunication technologies defined by a challenged environment. Section 3 presents a state-of-the-art network communication architecture. A case study on the application of a routing protocol based on Ant Colony Optimisation (ACO) algorithm within conventional network communication architecture is presented in Section 4. Section 5 presents a design methodology for ad hoc networks with a cross-layer architecture based on a multi-criteria decision making model designed to operate in disturbed environment and to guarantees temporal constraints. A conclusion and future directions are presented in Section 6.

II. CHALLENGED ENVIRONMENT OF AD HOC NETWORKS

Numerous studies investigating the challenging environment of communication networks focus on the behavior of their physical layer [5]. For this reason, a good understanding of physical layer will define suitable

 ISBN:978-1-61208-113-7

processing in network management and offer the possibility to avoid problems that directly relate to the network QoS.
Wireless networks have made remarkable advances and today, they seem to have unlimited capacities. Based on such advances in communication networks, a range of complex applications have been developed, many of which have significant QoS constraints. A realistic observation of networks environments and their physical layer performance makes it possible to detect inadequacies that lead to packet losses and to an increase in the transmission delay for wireless networks.

A. *Interference in the wireless environment*

It is important for wireless networks, especially for ad hoc networks, to define suitable access layer protocols that will minimize, if not avoid, packet collisions due to simultaneous transmissions. Ad hoc networks also require error protocols that counter frequent interference to radio transmissions and deal with the variable network topology of their network nodes. The very complex space geometry of the ad hoc network environment and various obstacles that may be encountered by the radio signals (buildings, bridges and tunnels, etc.) contribute to raising the level of interference. Various phenomena have a direct impact on wireless signal propagation. These include fading of the transmitted signal and multi-path propagation caused by physical phenomena such as refraction, diffraction and reflection. Research on these phenomena and their impact on network quality relies heavily on mathematical models used to represent signal propagation in a realistic manner [5][6][7].

B. *Review of propagation models for wireless networks*

Propagation models are used to simulate the attenuation of wireless signals in a particular environment. Generally, propagation models compute the power of the signal at the receiver as a function of the power of the signal at the transmitter. Depending on the features of a particular environment, one of three main propagation models may be used: free-space loss model, two-ray ground model and shadowing model. Most of the published results in the field of ad hoc wireless routing and broadcasting are based on free-space or two-ray ground propagation models, which are simplistic and idealistic. Indeed, these models are usually unable to capture the spatio-temporal variations of the signal power at the receiver. Therefore, a probabilistic model is more suitable for our study context designed to depict an environment that experiences dynamic events. The shadowing model defined in [8][9] can be used for more realistic propagation models.

The average large scale path loss for an arbitrary Transmitter-Receiver (T-R) separation is expressed in [8] as a function of distance by using a path loss exponent, n. The following equation expresses the average path loss PL (d) for a transmitter and receiver with separation d with d_0 as the reference distance.

$$PL(\mathrm{dB}) = PL(d_0) + 10\, n \log\left(\frac{d}{d_0}\right) + X_\sigma.$$

$X\sigma$ is a zero-mean Gaussian distributed random variable with standard deviation σ. An important feature of the shadowing model is its ability to simulate a wide range of environments in which fading and interferences are determined by simply adjusting the value of n.

III. BACKGROUND

In the context of wired networks, layered structures proved to be reliable for usage in numerous high speed communication technologies such as ATM (Asynchronous Transfer Mode) based on SONET (Synchronous Optical Network). Because of the behavior of their physical layer and the impact of the degradation of this behavior on the upper layers, wireless networks have entirely different requirements. Among studies carried out on performance problems in wireless and ad hoc networks, two categories of research activities can be identified. Advances on communication services in the context of TCP/IP communication architecture, is still considered for services innovations in addition to the new architectural design based on the concept of cross-layer.

A. *Advances in communication protocols based on TCP/IP*

Recent research activities on ad hoc networks have made important contributions to accessing the link layer. The result is the ability to minimize the number of collisions of the access link and to optimize the network resources as a spreading spectrum [10]. Other advances in protocols are in the domain of routing protocols. Various approaches, such as metaheuristics [11][12], have been taken to study the adaptive protocols for QoS optimisation.

The main difficulty of these communication architectures is the variation of delay between degradations in the physical layer and reactions in the upper layers.

B. *Architectural design of cross-layer communication*

Research activities on cross-layer architecture have been mainly focused on cognitive networks, which originate in cognitive radio [13]. Adaptation to changes represents an important topic for these networks. This category of research activities focuses on increasing the network performance. Literature on this research topic states potential advantages and direct impacts of the physical layer

 ISBN:978-1-61208-113-7

on operations of the nonadjacent high layers such as link layer, network layer and transport layer.
Researchers have attempted to face these important challenges. One of the proposed solutions is based on a cross-layer architecture, which identifies three interaction categories [14]:

- Direct communication between layers, based on variables of a layer visible to the others in runtime. Internal states of the layers have to be managed for this category.

- Sharing of a database between the layers: In addition to the shared database access, a research topic here is the design of interactions between the different layers.

- Elimination of stack structure: The components of communication architecture are autonomous. This category offers a great flexibility but represents a great challenge compared to the well-known organisation of protocols. Numerous research activities have studies cross-layer designs. In [15][16]. In cross-layer communication architectures, it is easier to define a network that has the knowledge of its environment. This is a characteristic that constitutes an important factor for ad hoc networks. In the meantime, a suitable architecture cannot be defined without the knowledge of related applications and their QoS. In Section IV, we present a case study based on the TCP/IP architecture that provides more information about network activities and their impact on the network QoS.

IV. CASE STUDY: ROUTING ALGORITHMS ON NETWORK BASED ON TCP/IP ARCHITECTURE

In this section, we discuss our case study of the application of metaheuristics inspired by the ACO (Ant Colony Optimisation) algorithm extension adapted to mobile ad hoc networks (AntHocNet) [17]. The case study addresses routing problems in a VANET (Vehicular Ad hoc Network) [18], a network that is gaining importance especially in the context of Intelligent Transportation Systems (ITS). VANET is an ad hoc network composed of vehicles with communication capacities and characterized by their mobility model. To address potential issues in the physical layer and the consequences for network performance, we employ the cross-layer concept. Before going into detail on our case study, we present an overview of VANET and the cross-layer concept.

A. Introduction to vehicular internetworking

Wireless networks comprise different categories of networks, such as the Mobile Ad hoc Network (MANET). MANETs are self-configuring ad hoc networks based on mobile routers and VANET (Vehicular Ad hoc Network). In a VANET network, communications may take place between vehicles (vehicle to vehicle) or between vehicles and roadside nodes. Important technological advances in VANET networks have led to the development of a variety of complex applications for ITS such as disaster management. Complex applications must meet QoS with a focus on temporal constraints to manage emergencies on the roads.

B. The concept of cross-layer architecture

The cross-layer concept is based on the principle of layered protocols which constitutes the foundation of the classical architecture of network communication. In the new generation of communication networks, namely cognitive and autonomous networks [14], communication is allowed between nonadjacent layers. This will give rise to a situation where changes on any particular layer can directly affect the quality and the operation of another layer of the hierarchy or affect various aspects of network management such as performance, faults and security. Thus, during the design of a network with cross-layer architecture and according to chosen interaction category, it is necessary to identify new relations between layers and specific processing requirements such as notification of events .

C. Routing protocol based on AntHocNet

AntHocNet is a hybrid algorithm that uses reactive and proactive mechanisms in order to discover routes. The AntHocNet algorithm defined in [17], was adopted in our case study. We also introduced a few modifications to address particular time constraint problems in complex applications. AntHocNet works in four phases: route setup, route maintenance, data routing and route repair.

- Route setup: At the beginning of each communication session between a source node and a destination node, the algorithm creates a special packet called *reactive-forward ant packet* which simulates an *exploration ant*. The *reactive-forward ant* packet is broadcasted at the source and along the network until it reaches the destination node. At the destination, the *forward ant packet* is discarded and a *backward ant packet* is created. The *backward ant* packet will follow the route taken by the *reactive-forward ant* in reverse and will set up a route to the destination at each intermediate route including the source node.

 ISBN:978-1-61208-113-7

- Route maintenance: The aim pursued in the phase of route maintenance is to either keep the existing routes or to find new routes to the destination. To do this, the source node periodically generates a special packet called *proactive-forward ant* packet which is transmitted to the destination through a random neighbour. The *proactive-forward ant* packet follows its route along the network until it reaches the destination. At the destination node, the *proactive-forward ant* is converted into a *backward ant* packet which, as in the phase of route setup will follow the reverse route and update the routing tables of the intermediate nodes.

- Data routing: The phases of route setup and route maintenance make it possible to find a set of routes to the destination. The phase of data routing consists of choosing one path from this set.

- Route repair: In AntHocNet, each node tries to maintain an updated view of its immediate neighbors at each moment, in order to detect link failures as quickly as possible and before they can lead to transmission errors and packet loss. The presence of a neighboring node can be confirmed when a hello message is received, or after any other successful interception or exchange of signals. The disappearance of a neighbor is assumed when such an event has not taken place for a certain amount of time or when a unicast transmission to this neighbor fails. When a failure is detected, the algorithm removes the responsible neighbor from the neighbor list that it maintains. Then, the algorithm checks the presence of a secondary route to the destinations. If no such route is detected, the algorithm informs its direct precursors by means of a special packet called a *link error packet* which contains all the unreachable. The same process is subsequently repeated until all the nodes in the networks are informed of the change.

The main changes we made in the AntHocNet routing protocol related to specific functions and the overall architecture of the algorithm. Our architecture is sequential and for the sake of simplicity, no concurrency is considered in the present implementation, called AntHocNet-1. Figure 1 illustrates the components of our routing algorithm.

Regarding specific functions, instead of the stochastic mechanism used in AntHocNet, we use a greedy forwarding mechanism in the data routing phase. In addition, while the authors of AntHocNet utilize a combination of delay and hop count as a pheromone amount, we use the total end-to-end delay from the current node to the destination in our implementation.

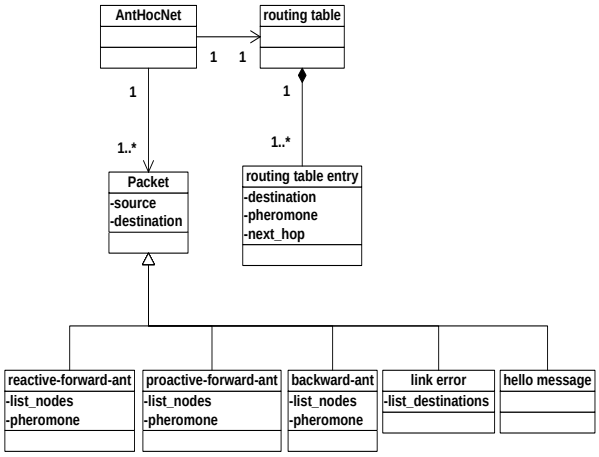


Figure 1: Structure of routing Algorithm AntHocNet-1

D. Experiments on VANET using a Routing Protocol

Our objective is to study the performance of the AntHocNet-1 algorithm in the context of a complex technology such as VANET by considering the temporal constraints as important criteria for assessing network performance. We also propose to study the behavior of this algorithm according to the network architecture by considering traditional (standard) structure TCP/IP with architecture dimensions which would be based on the cross-layer concept.

V. SIMULATION ENVIRONMENT

Experiment environment is made up of two simulators, SUMO (Simulator Urban Mobility), a microscopic road traffic simulator. In order to model the mobility realistically; we selected a tool called MOVE (Mobility model generator for VEhicular networks) [18]. The second simulator is ns-2, used for a communication network simulation [19]. Our experiment environment is illustrated in Figure 2.

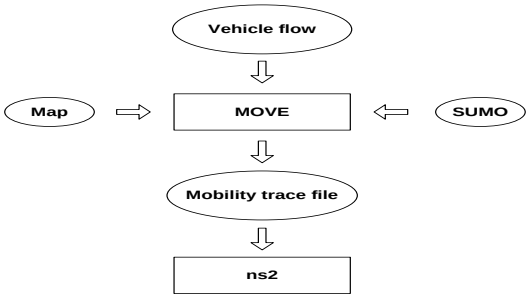


Figure 2: Experiment environment

In our experiment environment, we first identify a set of VANET network features including the topology, the Grid

MAP, vehicles in motion, etc. MOVE uses the SUMO output in form of mobility traces that realistically represent vehicle mobility according to the rules of the road such as traffic light coordination, signs, maximum allowed speed and priorities. MOVE communicates these mobility traces to ns-2. These traces will be used for simulation of VANET network such as the nodes mobility model.

A. Simulation Parameters

In our experiments, the geographical space for VANET is composed of a grid map representing a set of roads. The number of vehicles in our VANET is set to 80. Of these, 50 move according to a model generated by SUMO and the rest are generated randomly and are characterized by setting the periodic packet transmission to a rate of 2 CBR (Constant Bit Rate) packets per second.

The transmission range of the nodes is set to 400 m. We examine two communication protocols having an impact on the network performance: the MAC layer and the network layer. According to the ns-2 implementation based on the *IEEE 802.11/b* standard, the MAC layer is based on the protocol DCF (Distributed Coordination Function), while the physical and network layers will be instantiated. The propagation model for the physical layer is therefore the shadowing model since changes in its parameters will model environment changes and disruptions such as link loss. We use the AntHocNet-1 algorithm as a network protocol in order to study and analyze its performance when faced with environment changes represented by changes in propagation model parameters. We set the simulation duration to 200 s, then run each scenario 10 times. Table 1 shows our simulation parameters.

Simulation parameter	Value
Medium access protocol	DCF
Simulation time	200s
Shadowing model	n=variable, s =4.0
Transmission range	400 m
Transmission power	0.28 mW
CBR	2 packets/s, each packet length is 64 Ko
Number of vehicles	50
Maximum vehicle speed	50 km/h

Table 1. Simulation parameters

In these experiments, specific path losses are defined to represent a degree of environmental disruption. The path loss exponent parameter *n* varies from 2.0 (non-disturbed environment) to 2.1 (disturbed environment) according to the Shadowing model. We evaluate the behavior of AntHocNet-1 through two architectures: the classical architecture based on TCP/IP, in which communication is only between adjacent layers, and the cross-layer architecture, in which communications occur between non-adjacent layers. To test the concept of cross-layer architecture in our simulation, we established the following process: The physical layer monitors, by measurements, the wireless environment in order to detect disruptions. If a disruption, evidenced by changes in the path loss exponent, is detected, the physical layer directly notifies the network layer. In this study, we simulated physical and routing layers interactions by temporal notifications. The AntHocNet-1 protocol reacts to the notifications by executing the phase of route repair. In this experiment we change the frequency of disturbances from every 10 s to every 100 s, while keeping the length of each disturbance constant set to a value equal to 5 s.

Simulation results in terms of end-to-end delay are illustrated in Figure 3.

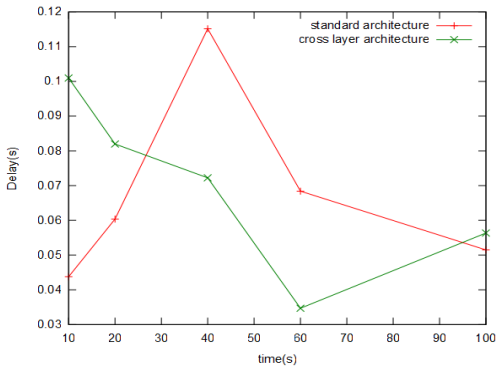


Figure 3: Simulated cross-layer vs. standard architecture behaviors

A. Towards a cross-layer network architecture based on multi-criteria decision making

The results show the impacts of disturbances in the physical layer on the performance of the routing protocol. The tested network uses the traditional communication architecture set up according to the TCP/IP model. It is worth noting that this architecture is based on requests between layers and the delays for reactions especially in a congested network, a situation that causes a loss of communication packets.

In order to support complex communication requirements of applications, the network communication architecture proposed in this study is based on the concept of cross-layer. This section analyzes the consequences of potential side effects caused by significant changes in the frequencies of physical layer.

 ISBN:978-1-61208-113-7

B. Cross-layer operating scenarios

The subject of our study has been QoS requirements of complex applications. These requirements can be multicriteria and one of the criteria can consist of real-time constraints. We hence stress the importance of communication between the layers through notifications followed by real-time reactions. Thus, for mobile ad hoc network, significant changes of channel radio must be transmitted immediately after their occurrence to the concerned layers .

According to our previous works [1][20], an increase in the frequencies of environment disturbances events will lead to performance degradations defining networks congestion or systems saturation. These disturbances manifest themselves by a failure to process events within the temporal constraints. The same behaviour is then likely to occur in a mobile ad hoc network tested in this study and illustrated in Figure 3. The network based on cross-layer concept responds to this condition by continuous reconfigurations of its network protocol presented by route maintenance in AntHocNet-1 algorithm and which seems to lead to performance degradation according to disturbances frequency.

C. Methodology of a dynamically reconfigurable architecture

The unpredictable environment behaviour in mobile ad hoc network and the risks of congestions and saturation described previously, make it difficult, if not impossible, to validate the design of the cross-layer architecture in order to guarantee the required QoS. This is even more critical when applications have to deal with dynamic criteria that represent their QoS. Thus, cross-layer architectures presented in section III cannot be used systematically in the context of our study despite their performances observed for some physical layer states compared to the standard communication protocol. We define a methodology of an adaptive architecture design that is more suitable for environment changes. An adaptive architecture faces the problem of experimental identification of parameters for the communications between layers. Thus, parameters such as waiting delays between the occurrence of the events and the reactions of higher layers will be adjusted to the QoS requirements of applications.

We consider learning by reinforcement as a direction to readjust the parameters of communications. This will be possible by using the network management system feedback on the global quality of the network in a given state of the environment and by the usage of parameters for a given architecture configuration. A decision-making process will also be integrated within the reinforcement learning in order to evaluate the network QoS based on a multicriteria objective function to optimise. Finally, this study will consider numerous areas of applications to validate the communication architecture. It also will study its adaptations and the associated dynamic decisions based on different criteria defined by applications requirements.

V. CONCLUSION AND FUTURE DIRECTIONS

This paper presented a realistic context of the application of ad hoc networks and the new requirements for real-time applications. It has presented experiments on a routing protocol based on the Ant Colony Optimisation (ACO) algorithm alternative on a network traditional architecture used for TCP/IP model and a cross-layer concept. The results show performance degradations according to the physical layer perturbations for each architecture model.

Study of the cross-layer concept demonstrates potential advantages of these networks, especially in presence of temporal constraints but not in the case of extreme disturbances. The required architectural adaptations may lead to a cognitive network that represents a promising solution for today's applications, which are mainly based on ad hoc networks. Consequently, suitable network communication architecture is defined by its ability to adapt its parameters and configuration to the changes of the environment network, especially following the notifications of metrics related to the quality of the radio channel along the routes.

This work enabled us to highlight a new direction for the communication architecture of cognitive vehicular networks operating under disturbed environment such as the quality degradation of the radio channel along the routes. This direction consists of considering a dynamic reconfiguration of the communication architecture. Consequently, suitable network communication architecture is defined by its ability to adapt its parameters and configuration either by the traditional architecture or the cross layer architecture based on autonomous components. Reinforcement learning is a suitable approach to identifying on line the best architecture among the traditional TCP/IP architecture that feature waiting times in communication between layers and cross-layer architecture based on non-adjacent layer events notifications.

Validating cross-layer architecture in the context of this study requires realistic applications considering their QoS requirements. Applications for network resource management will be considered for the validation of future approaches. We will also consider simulations based on autonomic components in our future study to examine the communication network architecture based on the cross-layer concept.

ACKNOWLEDGMENTS

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC)

REFERENCES

[1] I. Benyahia and D. Lapointe, "A Complex Applications Framework Supporting Adaptive Routing Strategy for Ad hoc Networks", Advanced International Conference onTelecommunications AICT'06, pp. 74-80, 2006.

[2] P. Golding, "Next Generation Wireless Applications: Creating Mobile Applications in a Web 2.0 and Mobile 2.0 World", by: Paul Golding, Wiley; 2 edition, 2008.

[3] I. Chlamtac, M. Conti and J. J.N. Liu , "Mobile ad hoc networking: imperatives and challenges", Ad Hoc Networks, Elsevier, pp 13-64, 2003.

[4] A. Willig, "Recent and Emerging Topics in Wireless Industrial Communications: A Selection", IEEE Transactions On Industrial Informatics, Vol. 4, No 2, pp 102-124, 2008.

[5] J. M. Dricot and P. De Doncker, "High-accuracy physical layer model for wireless network simulations in NS-2", International Workshop on Wireless Ad hoc Networks (IWWAN), 5 pp, 2004.

[6] J. M. Dricot, P. De Doncker, E. Zimanyi and Fr. Grenez, "Impact of the Physical Layer on the Performance of Indoor Wireless Networks", in Proceedings of the International Conference on Software, Telecommunications and Computer Networks, October, pp 872-876, 2003.

[7] Y. Yu and S. L. Miller, "A Four-State Markov Frame Error Model for the Wireless Physical Layer", IEEE Wireless Communications & Networking Conferences, (WCNC), pp 2053 – 2057, 2007.

[8] R. Akl, D. Tummala, and X. Li, "Indoor Propagation Modeling At 2.4 Ghz For IEEE 802.11 Networks", Proceedings of WNET 2006: Wireless Networks and Emerging Technologies, paper no. 510-014, 6 pgs. July 2006.

[9] N. Patwari and P. Agrawal, **"**NESH: A Joint Shadowing Model For Links In a Multi-hop Network", IEEE International Conference on Acoustics, Speech and Signal Processing, pp 2873 - 2876, 2008.

[10] Q. Li, and A.Q. Hu, **"**A high efficiency spread spectrum modulation based on Non-Strictly Orthogonal Sequences", IEEE International conference on Wireless Communications & Signal Processing, WCSP, pp 1 - 3, 2009.

[11] D. Montana and J. Redi, **"**Optimizing Parameters of a Mobile Ad Hoc Network Protocol with a Genetic Algorithm", GECCO'05, The Genetic and Evolutionary Computation Conference), pp 1993-1998, *2005.*

[12] I. Bouazizi**,** "ARA - The Ant-Colony Based Routing Algorithm for MANETs", Proceedings of the International Conference on Parallel Processing Workshop, pp 79-85, 2002.

[13] S. Haykin, "Cognitive radio:brain –empowered wireless communications", IEEE Journal on Selected Areas in Communications, vol, 23, pp 201 – 220, 2005.

[14] H. M. Qusay, "Cognitive Networks: : Towards Self-Aware Networks", John Wiley, 2007.

[15] S. Sharkkottai, T.S. Rappaport and P.C. Karlsson, "Cross layer design for wireless networks", IEEE Communication Magazine, pp 74 – 80, 2003.

[16] V. Srivasta and M. Motani,"Cross-Layer Design : a Survey and a Road Ahead", IEEE Communication Magazine, pp 112-119**,** 2005.

[17] G. Di Caro, F. Ducatelle and L. M. Gambardella, "An Adaptive Nature-inspired Algorithm for Routing in Mobile ad hoc Networks", European Transactions on Telecommunications, pp 443–455, 2005.

[18] F. K. Karnadi, M. Z. H. Mo and L. K. Lan, "Rapid Generation of Realistic Mobility Models for VANET", IEEE Wireless Communications and Networking Conference, pp 2506 – 2511, 2007.

[19] http://www.isi.edu/nsnam/ns/ - 17.11.2010.

[20] I. Benyahia and V. Talbot, "Optimizing the Architecture of Adaptive Complex Applications Using Genetic Programming". Proceedings of the Fourteenth International Conference on Distributed Multimedia Systems, pp. 27-31, 2008.

 ISBN:978-1-61208-113-7

Regional City Council mGovernment Case Study: Success Factors for Acceptance and Trust

Shadi Al-Khamayseh
Higher College of Technology
Dubai Women's College - Dubai City, UAE
shadi.alkhamayseh@gmail.com

Elaine Lawrence
School of Computing and Communications
University of Technology Sydney, Sydney, Australia
Elaine.Lawrence@uts.edu.au

Abstract— This paper is the second of a series of investigations into the key success factors for mobile government services (mGovernment) in Australia, Jordan and the United Arab Republic. The case study in this paper concerns a large regional city council in Queensland, Australia. It provides valuable insights into the adoption of mobile government services and applications over the council's wired and wireless network. Here, we report on a major achievement of the study namely, the identification of success factors for mobile government.

Keywords-mobile; government, success factors; case study

I INTRODUCTION

This case study outlines the mobile government systems implemented at a regional city council in the State of Queensland, Australia. The rationale behind the selection of local government is the high number of community issues with which local governments must deal compared with other government levels (i.e., state and federal). These community issues form a pressure on the local governments to improve their services through adopting the latest mobile and wireless technologies. The analysis of the case study generated a number of success factors that should be considered when planning for, and while implementing, the mobile government systems and services. This paper is part of our study of mobile government services [1] [2]. Here we report on one major case study from an analysis of eight (8) other case studies from Australia, the United Arab Emirates and Jordan. Section 2 provides the background to the case study and Section 3 outlines the methodology. Section 4 reports on the case study and the conclusions are found in Section 5. A table of success factors is set out in Appendix A.

II. BACKGROUND

For ethical reasons the information is de-identified. Thus, the paper refers to this regional local government city council as Queensland Regional City Council2 (QLDCC2) in Australia. The team manager of technical services was interviewed for this case study. The goals of this study include an exploration of the (a) success factors and (b) economic aspects of the acquisition of wireless and mobile technologies in the public sector. The objectives deriving from those goals are:

TABLE 1. OBJECTIVES & QUESTIONS

OBJECTIVES	**QUESTIONS**
An assessment of the categories of mobile and wireless technologies use in public sector.	*What patterns of acquisition emerge from the current wireless and mobile technologies and the perceived needs for mobile government?*
The establishment of a basis for understanding the current and future success aspects of mobile government.	*What characteristics of the categories of mobile government services contribute to the patterns of acquisition*?
An evaluation of the mobile government adoption issues, including managerial issues and the centralization and/or decentralization of decision making.	*What issues arise from the rapid acquisition of mobile and wireless technologies and how important have those technologies become to the organization?*
The establishment of a basis for understanding the current and future economic aspects of mobile government.	*How will the organization balance the need for technological changes with the need to continue the accomplishment of routine tasks?*

These questions were used for all eight case studies and the answers were recorded and transcribed and later analyzed using nVivo, a qualitative data analysis software package.

III. METHODOLOGY

The above questions along with the main research question: "*What are the factors that contribute to successful mobile government? How? Why?"* reinforce the exploratory nature for this research [3]. The author [4] defines the methodology as "*the strategy, plan of action, process or design lying behind the choice and use of particular methods and linking the choice and use of methods to the desired outcomes*". According to [5] "*Case*

 ISBN:978-1-61208-113-7

study can be seen to satisfy the three tenets of the qualitative method: describing, understanding, and explaining". This forms the rationale behind selecting case studies as the research strategy. The research question type also led to choosing the case study methodology. The core research question includes three categories of the enquiry questions series: 'What', 'How', and 'Why' [3]. The 'What' part here aims to develop propositions for further inquiry. The 'How' and 'Why' parts "are more explanatory and likely to lead to the use of case studies" [3]. In [6], a case study is defined as:

> *an in-depth investigation of a discrete entity (which may be a single setting, subject, collection or event) on the assumption that it is possible to derive knowledge of the wider phenomenon from intensive investigation of a specific instance or case.*

Tellis [5] acknowledges that case study methodology has been used extensively in many fields such as: "*law, medicine, government and in evaluative situations"*. He points out that "*The government studies were carried out to determine whether particular programs were efficient…*" [5]. In this case study, QLDCC2 is a regional council that serves an area of 3,127 km^2 and a population of 300,000. The council has approximately *'1300 employees'* and *'has hundreds of different types of services .., from libraries to cemeteries, to water operations, inspectors, building inspectors, so there's a whole raft of different business groups out there, … and … different business needs'*. The council utilizes *'40 Blackberry handsets*' as it *'needed to provide some secure service for managers so they could get access to emails and their calendar, while they were out in the field'*. The council also uses the NextG wireless network so that the mobile workforce can communicate with the council's systems from outside the council offices. Furthermore, the technical manager explained that *'it's actually not just our mobile users that use NextG; we have about 5 or 6 locations, where we use NextG. One of them is the cemetery up the road; they use NextG, because they are out of the range for ADSL - they can't get the high speed ADSL, so we can provide high speed connectivity back into the network through the wireless NextG*".

QLDCC2 has also implemented Wi-Fi hotspots in their libraries. The manager noted that in their *'main libraries we provide free Internet access'*. Recognizing its diverse community, the council arrived at its 'Mobile Libraries' initiative - '*basically it's a bus*, a *mobile library bus that drives around various sites'*. At first, the mobile library used to *'go to fixed stops, where there was a phone line into which they could plug'*. Then, the council realized the benefit of the available wireless technologies, so *'now that we've extended it to be fully mobile, it uses the Telstra NextG service'*. The manager explained that *'it's cheaper doing that than it was previously having a fixed line at those 25 sites, for which we were paying rentals, and you know you go from providing a 64k service through to a 7.2 megabit service, so a huge increase in actual performance'*.

The mobile library *'visits areas like, outside retirement villages, schools, sort of locations where there … isn't easy access to the fixed library'*. The mobile library houses items covering a wide range of formats, including books, magazines, DVDs, music CDs, toys, large print and audio books - something for all ages. The manager explained that the mobile library *'provides Internet access for the public when the library turns up and also it provides some online services for the library staff that work in that library, so they can check out books, record books that get loaned out and come back in, and it also provides several Personal Computers in the bus, for the public to come and do online catalogue queries, and surf the Internet'*. In terms of the benefits achieved, the manager explained that *'now in words of ability, it can now go to anywhere within our region and it's got full network coverage through Telstra, and we can provide the service to the public for Internet, online catalogue searching and also to our staff that actually man that library'*. Figure 1 depicts the mobile library.

Figure 1. Regional Mobile Library QLDCC2 2008,
(Source: Welcome to QLDCC2 libraries, viewed 23/10/ 2008 http://library.QLDCC2.qld.gov.au/sitePage.cfm?code=mobilelibraries).

IV THE CASE STUDY

The manager explained that the main driver behind implementing the mobile government system was the fact that *'the technologies are out there'*, but this did not mean that all business units within a regional city council should utilize the technology. Investing in technology should bring benefits and/or profits to the business unit to justify the need to invest -*'we don't necessarily implement technology just for the sake of implementing it'*.

The business case perspective was highlighted as an important factor by the manager from QLDCC2. Investing in technology should bring benefits and/or profits to the business unit and possibly to the organization as a whole. It is not necessary that the business unit is aware of, or understands, the new technologies, as it is the IT division's responsibility to *'liaise a lot with the business to find out what their requirements are and make them aware of these technology changes'*. The manager stated *'we* [IT division] *work with the business to ascertain what their requirements*

are and what benefits they'll get if we implement the technology'. This highlights the importance of the liaison among the IT division and all business units. He also cautioned that *'there are costs and overheads, as we implement the technology, you are going to need more staff to maintain your system, there's a cost to implement it, so it's done on a case by case basis'*. Determining the costs of implementing mobile government systems is a joint effort where the IT division provides its experience too. For example, the manager acknowledged that *'security costs were a big one'* and these costs are determined by the IT division.

To enhance cooperation, the manager emphasized the importance of having a *'good relationship with the business'* through maintaining open communication channels between the IT division and all business groups, so if *'they require new technology solutions, they might come to us… and we work with them to find a solution'*. To sum up, the case study showed that implementing mobile government systems through investing in mobile and wireless technologies is based on business case proposals which require:

- IT liaison with the business units to understand their requirements;
- IT liaison with the business units to make them aware of the available technologies that meet their requirements;
- A valid and justified business case that supports the decision to invest in the technology.

Forty (*40) Blackberry handsets* were introduced *'about 3 to 4 years ago'*. To do so, the council *'looked at a few different technologies, and found Blackberry was … the most advanced at that time. Blackberries provided a good comprehensive solution, that was secure and you could also manage the devices that you had from the central system'*. The council has been utilizing different mobile devices distributed to many different users, such as the *'standard normal size laptops used by the water operations'*, *'Blackberries used by the councillors and managers'* and *'touch screen tablets for those who need more organizer devices'*. The manager justified utilizing the different devices due to the fact that *'the different business areas are defining requirements that they need'*. This highlights that the requirements play a role in mobile device selection. For example, a business unit might need *'access to either online data when they're actually in the field doing GPS measurements or whatever … so they can sign off jobs, out in the field'* and *'for others it might be to just record static information as they are moving around, now those guys … don't have live connections to the network because they don't need it'*.

Although some business units do not have a business need that requires live connection in the field, they still use the mobile technologies (i.e. mobile devices) to collect static data such as *'GPS locations'* of any asset *'that needs fixing'*, and *'when they* [the users] *come back to the office they just download the information, it goes straight into the system'*. The manager illustrated that utilizing the wireless technologies through adding the live connection depends on *'what the requirements were and how they actually work'*. The council has to consider *'weighing up the costs to provide that service to the benefits'* gained. The manager reported that the *'live connectivity'* facility used by the *'plumbing inspectors'* and *'operations'* staff costs the council about a hundred dollars per staff member per month, but resulted in many *'tangible benefits'* such as saving *'a lot more down the track'*. The manager demonstrated this through the following example:

A good example is the plumbers, ..., going out collecting data, but they'd spend several hours, and they'd come back to the office, updating all the systems typing it all in, so they benefited in a big deal by … simply having a laptop, NextG connected out in the car and record the information as they do their job, and saves them travelling, gets synchronized straight away, it's instantaneous, so it's live on the system, it saves them having to spend several hours in the morning or afternoon coming back in the office, to get Internet information etc.

Another factor that plays a role in achieving a successful solution is the access devices. In this case, the selection of Blackberries was partly based on their ability to house new applications for any future emerging business need. The manager illustrated that *'the primary use, initially for the Blackberries was to provide email and calendar access for councillors and managers'*, but *'the Blackberry platform we chose because you can actually have applications written, so you can interface with the systems* [corporate systems]*'*, so *'if the business need was there, to actually provide some functionality* [i.e. to access the corporate systems] *that way, then we could deliver it that way'*.

The implementation of the mobile government systems in the council resulted in utilizing *'over a hundred laptops, which are now configured for that service and they've taken that up in about the last year, so there's been a real explosive growth and that goes through from directors, to managers and to field operation staff'*. The manager also highlighted the speed of data transfer over the network as an important success factor for the mobile government system. He acknowledged *'certainly the speed that we can get through that system just makes the whole thing much more usable'*.

Finally, the manager outlined the features of the mobile government system the council aims to achieve. He explained that: *'we're trying to implement a system that doesn't add a large overhead in the back office systems, and a system that doesn't need specialized tailoring for every application to make it sort of, device dependent, in that we're trying to … implement a system that can operate generically on a range of devices, and that we can support cost effectively'*.

Mobile government has four main constituents (Government (G), Employees (E), Citizens (C) and

Business (B). In this case study, the council provides a number of G2C m-Government services or applications. The manager explained that in order to develop such projects it is important to *'have the figures'* as they demonstrate that *'there's another service that could be done'*. For example, such figures could show *'how pervasive … 3G web surfing technology is with the public'* which supports decision making on whether to provide it. The manager also highlighted that for a mobile project to succeed, it should mimic the real process as much as possible. For example, the mobile library solution: *'it's a free service to the public'* and the council has *'extended it now so they can provide Internet access'* because in the *'main libraries we provide free Internet access, to the public and this mobile library is now doing the same service'*.

Although a high mobile and wireless technology penetration amongst citizens results in more success for G2C m-Government services, the manager noted that it *'is not the local council's role'* to market the mobile and wireless technologies to them, because the *'the council doesn't control the wireless market.* According to the manager, such G2C m-Government services *'could probably be provided quite easily'*, but deciding to develop them *'comes down to the penetration of the market'*. This highlights the importance of knowing the penetration figures and discovering citizens' needs according to these figures. To sum up, the manager has recommended the following success factors for G2C m-Government services:

- Broadband availability
- Attain representative figures of mobile usage
- Broadband coverage over the network
- Easiness and readability of services on mobile devices
- Mimic the real process which people know and understand

The manager focused on the G2E m-Government services, saying that the *'operations guy, who are in vans, have laptops to get back into the council's systems. ˜Now all those particular users are provided with a laptop with a NextG card, which is configured with a corporate service, so it comes corporately back into our network'*. The manager illustrated that the implementation of mobile government systems, especially the G2E m-Government services, has not affected jobs in terms of loss or growth. It has only resulted in benefits, such as increased productivity of the workforce. In order to successfully implement the system, it is important to understand the users' skills, especially as some users (e.g., plumbers) *'never used the computer before'* and after implementing the mobile government system, they will suddenly be dealing with a mobile device that might well be connected to the corporate systems. Managing the transition from the old style of writing on paper to the most up-to-date technologies creates a challenge to the success of the system. The manager acknowledged the challenge and recommended that one should *'go through a training issue with the staff, they have to be trained to get the benefit from it'*.

The manager emphasized that *'there is a great deal of learning in there'*, and concluded that it is essential to make changes to how things work. The manager also highlighted *'getting that **acceptance** and **trust'*** as critical to achieving a successful mobile government system. He suggested that *'some of them* [mobile workforce] *are even thinking, 'oh well now they can track us with this gear and know what we're doing'*. To gain that trust is *'mainly just working through that issue with the people'*, to make them aware of the reason behind implementing the system, and its benefits. The manager recommended addressing their concerns and answering *'no we're not doing it so we can keep track of you; it's been done so you can be more effective'*, and make them aware that *'it's not because we want to watch exactly what you're doing from a time management thing, it's more from a resource management of where to deploy the resources'*. Thus, gaining acceptance and trust is *'something they have got to manage really carefully'*.

To sum up, achieving a G2E service depends on the following success factors:

- G2E m-Government solutions must be configured with a corporate service
- Staff trust
- Staff acceptance
- Training
- Change management.

An important issue that the manager raised was the selection of providers. The chosen network plays a role in ensuring the security of the systems. The council has chosen Telstra's NextG network for their mobile services, such as the mobile library. Telstra was chosen for many reasons, but one reason was the *'GWIP* [Government Wideband Internet Protocol]" which is *'part of the Telstra Next Generation IP platform'*. The manager explained that equipping the *'operations guys'* with *'laptops with NextG cards'* started *'in* 2007 *as Telstra rolled out their NextG service'*. The manager explained that *'prior to that we did tests with their* [Telstra's] *old CDMA network and GPS - GSV network, but the speeds weren't really there'*. Here, the manager highlighted the speed factor as important *'to effectively deliver a corporate application out in the field; now from NextG, we can get good speeds'*. This highlights the importance of the available speeds offered by the provider's network.

The manager emphasized that the provider's infrastructure is a core selection criterion as it plays a role in the security of the systems. The manager justified his department's decision to go on with NextG from Telstra rather than 3G from other providers as *'going on with Next G …* [from] *Telstra, we can have that secure VPN that goes directly into our network, and it removes access through the Internet.'*

In summary, it is essential to enter a long term partnership with a reliable provider to achieve a successful

 ISBN:978-1-61208-113-7

mobile government system. It is necessary to understand the following providers' success factors to help in selecting the right provider:

- Ability to secure an access path to council systems
- Data transfer Speed
- Provider's infrastructure and its compatibility with council systems
- Provider's network speed
- Provider's future plans.

The manager highlighted that catering for mobility might require updating current systems to ensure security. The council has *'spent the last, about two years in going through planning for a network upgrade that can cater for mobility … in a secure way'*.

In addition to the careful planning required to update the department's current network, the manager discussed the importance of identifying and understanding the risks, threats and vulnerabilities that mobile government systems are subject to, such as *'viruses'*, *'spyware'*, *'Trojan Horses'* and *'worms'* that could find their way to the system when users access the Internet using council's mobile devices (i.e. PDAs and laptops). As operational countermeasures, the council does not *'allow connection on Internet to those unless it goes through our firewalls, and our Internet servers, so I can't take my laptop home, plug it into my Internet service provider at home, it doesn't work, we block it off, it's a security halt, because once that goes in, then you know, viruses and spyware and all that stuff can get on the machine'*.

The manager stated that the major security issues arise when *'going back into the council network'*. For this, the mobile library solution incorporates *'two servers'*. To ensure security, *'staff PCs and the router, that's got a corporate card off, so that basically goes into the Telstra cloud and securely goes into a VPN* [Virtual Private Network], *that pumps it directly back into our corporate network and it's coded so you can only do it with that particular SIM* [Subscriber Identity Module]. *You need a SIM code as well to put through and there is authentication all the way through* [this means the user must already be logged into 'Windows' to actually get connected to the network], *so it's fairly secure'*. The coding and authentication procedures ensure security of the data in case of device loss or theft.

In this case, the solution ensures security by providing separate paths for corporate data and Internet traffic. The public get direct access to the Internet, but staff can seamlessly interact and quickly transmit vital data through a dedicated path so that *'it doesn't go on the public Internet at all, it simply goes and hits the VPN gateway in the Telstra network and pumps directly back into our network. Now, for it to come in on this side we've got a Government Wideband Internet Protocol (GWIP) 20 meg service, and that takes in all those NextG cards and also some of our smaller remote sites that use GWIP"*.

One of the main threats is the loss or theft of the access devices. The manager demonstrated using the following example:

> *If the device gets stolen that might have a NextG card connected to it, simply turning it on wouldn't get them connection into the network. Every time the user turns on the laptop, they've got to logon as a local user on the laptop to start with, after they get through that security level, there is then a 4 digit PIN code which they must enter every time that the NextG device connects up to the network, ,,, they then have to logon to the CITRIX desktop, and that's the normal desktop or network logon, which they then have to enter, so there's like three levels they go through. Also, the actual service is keyed to that SIM card, so if they didn't have that SIM card they couldn't connect into the network.*

To sum up, to achieve a successful mobile government system, it is crucial to consider the following success factors:

- Update the current networking solutions for securely catering for mobility
- Awareness of all the risks, threats, and vulnerabilities to mobile government systems
- Use as many security levels as possible to ensure the ultimate security of the access devices and the systems
- Follow password selection best practice to generate passwords
- No sensitive information to be kept on an access device.

The manager commented that the state government so far has not pushed local councils to implement mobile government systems, but he highlighted that the government has a crucial role to play in helping them succeed in implementing the system. There are many *'ways they could help'*; one way is *'to bring the cost down and to bring the saturation even further'*. The manager argued that *'as the market gets saturated more and more uptake comes, and then you get more competitors, such as Optus rolling out their wireless networks, as that gets rolled out, that raises competition'*. This competition has already resulted in a drop in prices; for example, *'we pay a certain amount with Telstra now, for provision of the service, a couple of years ago it was more expensive, but now it's getting cheaper'*. The government could also negotiate with service providers to get the best deals for its departments. The manager acknowledged the importance of such negotiation; one result was that *'we use the state government contract for mobile phone contracts'*. The government negotiations with providers resulted in the creation of the *'Telstra Next IP Network, and it's basically made up of a couple of products like NextG, GWIP, Ethernet Campus, and it's all about providing … faster data services and remotely integrated data services on a big platform'*.

 ISBN:978-1-61208-113-7

V CONCLUSION

This case study highlighted a number of the potential benefits of implementing mobile government. According to the manager, *'there's a number of benefits'* for the mobile work force such as *'the exact recording of information'*, and *'they can actually work quicker out there, it becomes a time saver for them'*. The manager also identified a number of benefits in using the wireless network to connect some of the fixed locations into the council's network. This is usually needed when high speed ADSL is not available in the location, and the benefit is in *'providing high speed connectivity back into the network'*. The council found earned benefits in *'converting it* [locations] *from a fixed line, or a fixed ADSL or a fixed ISDN line that goes to the site, to a mobile NextG connection'*, and the benefit was that *'we've actually saved a lot of money … you change your cost from maybe three to five hundred dollars a month, down to a hundred dollars a month, so you get a doubling in speed or tripling in speed'*. Implementing G2E m-Government services has benefited the council in many ways, such as *'saving overtime costs'* as a result of no longer *'entering in information manually, which now they can enter out in the field'*. The manager affirmed that it *'increases productivity, they* [mobile workforce] *can actually get more work done or provide more services to the public, or respond quicker to the public, because they can get information online out there, like job dispatching. The guys have laptops always on out there, so when the jobs come through, they can see the details straight away when they're out there in the field'*. This shows that the return on investment not only benefits the government council, but all its constituents, such as the employees and citizens. To sum up, to implement mobile government in any council it is important to understand the potential benefits to be gained from such a system in order to justify the costs and to assist in the decision making process.

The analysis of the QLDCC2 case study has revealed a number of mobile government success factors. Table 1 in Appendix A summarizes the different success categories and their success factors.

REFERENCES

[1]. Al-khamayseh, S. and Lawrence, E. 2005, 'Mobile government – converging technologies and transition strategies', 4th International Conference on Electronic Government Trauner, Copenhagen, Denmark, pp. 358-365.

[2.] Al-khamayseh, S. and Lawrence, E. 2006, 'M-government success factors: a roadmap for developing interactive mobile government', 5th International Conference on Electronic Government within DEXA 2006, Krakow, Poland.

[3] Yin, R.K. 1994, Case study research: Design and methods, 2nd edn, Sage Publishing, Beverly Hills, CA.

[4]Crotty, M. 1998, The foundations of social research: meaning and perspective in the research process, Allen & Unwin, Sydney, Australia.

[5] Tellis, W. 1997, 'Introduction to Case Study', The Qualitative Report, vol. 3, no. 2. http://www.nova.edu/ssss/QR/QR3-3/tellis2.html, Last viewed 26 May 2010

[6] Gorman, G.E. and Clayton, P. 2005, Qualitative research for the information professional : a practical handbook, Facet Publishing, London, UK.7.

APPENDIX A

Success Factor(s)	QLDCC2 Case Study
Drivers	• The availability of technology
Solutions	• Easy management • Up-to-date • Selecting of access device • Data transfer speed • One platform • Device independent solution • Low overhead on the back office systems
Constituents (Citizens)	• Broadband availability • Mobile penetration and usage • Easiness and readability of (G2C) services • (G2C) services to mimic the real life process
Constituents (Employee)	• Staff trust • Staff acceptance • Training • Change management
Providers	• Security • Data transfer speeds • Infrastructure • Compatibility with the department systems • Future plans
Business Case	• IT liaison with business groups • Understanding business requirements • Awareness of available technologies • Justifying the needs and costs • Reflect on benefits
Security	• Update department's network • Awareness of all the risk, threats and vulnerabilities • Implement as many security levels as required • Best password selection procedures • No information to be kept on devices
Government Help	• Negotiation with providers
Benefits	• Understanding the benefits

 ISBN:978-1-61208-113-7

Evaluation of Buffer Size for Middleware using Multiple Interface in Wireless Communication

Etsuko Miyazaki
Ochamonizu Univercity
2-1-1 Ohtsuka, Bunkyo-ku 112-8610
Tokyo, Japan
Email: etsuko@ogl.is.ocha.ac.jp

Masato Oguchi
Ochamonizu Univercity
2-1-1 Ohtsuka, Bunkyo-ku 112-8610
Tokyo, Japan
Email: oguchi@computer.org

***Abstract*—Although a variety of wireless interfaces are available on mobile devices, they still provide only low throughput so far. When coverage areas of these different technologies overlap, mobile devices with multiple interfaces can use them simultaneously by mechanism of Bandwidth Aggregation. However, there are some performance problems for Bandwidth Aggregation on Network Layer and lower Layer which derive from TCP congestion control mechanism. We have proposed advanced Bandwidth Aggregation on Middleware for the purpose of avoiding there problems. In this paper, we have evaluated buffer size for receive-side Middleware.**

***Keywords*-component; Multiple interface; Middleware; Buffer Size; IEEE 802.11**

I. INTRODUCTION

The growth of mobile Internet communication stimulate developments of a variety of wireless technologies: for example IEEE 802.11, Bluetooth and WiMAX. Although some of them have relatively broad bandwidth, they still have lower throughput than wired connection such as Ethernet, and are able to be accessed only in limited areas. It is possible to have more efficient mobile Internet service using multiple interfaces simultaneously, when we are in areas covered by several services of wireless technologies. Bandwidth Aggregation which use multiple interface simultaneously is proposed as advanced way to access Internet from mobile node.

Among several research works, seamless vertical handoff from one interface to another has been addressed[1]. However, we have not achieved Bandwidth Aggregation in practical use. Those technologies give us better mobility support, reliability and resource sharing. Thus, we have proposed and evaluated an innovative mechanism of Bandwidth Aggregation in this paper.

II. BACKGROUND OF THIS RESEARCH WORK

A. Bandwidth Aggregation in Various Layer

Bandwidth Aggregation is supposed to be realized on several layers, while they have merits and demerits respectively.

An approach on Datalink layer[2] will give most effective result, will give the most effective result, and upper layer do not need to care about Bandwidth Aggregation. However, we can install it only world using same protocol for datalink layer and have to install specific hardware to their nodes.

An implementation in Network layer will provide efficient Bandwidth Aggregation by intelligent methods. The advantages using Network layer are they perform transparently to widely used Transport protocol such as TCP and UDP. However, TCP may not achieve estimated efficiency due to a possibility that they receive packets in incorrect order. These problems cause congestion control more than required.

In Transport layer, they have congestion window for each path. It enables more effective transport by doing packet distribution and retransmission for each path[3]. However, the system has to be installed into each operation system in all the end-end way.

An implementation on Application layer does not demand to replace current operating systems[4]. However, there are variety of applications and it is difficult to implement aggregation method for all of them. After connections established, we have to consider how to distribute packets for each connection.

B. Packet Loss Problem in Bandwitdth Aggredation on Network Layer

If multiple interfaces are used for concurrent communications, there are possibilities that receiving node may take packets incorrect order. In such a case, receiver recognizes occurring of packet loss incorrectly due to receiving packets different from expected order of packets. Then TCP requests retransmission unnecessarily. This is one of problems in Bandwidth Aggregation on Network Layer.

For the purpose of eliminating this problem, Earliest Delivery Path First (EDPF) was proposed[5]. EDPF is implemented to the node in which path is separated from sender to receiver. EDPF chooses on which path each packet should be sent in consideration of their bandwidth, delay and congestion. EDPF decides the fastest path to transmit the packet to receiver node. All packets are sent by the route on which estimated time is the shortest. Therefore, receiver can receive any packets in correct order. It makes Bandwidth

 ISBN:978-1-61208-113-7

Aggregation effective as estimated efficiency in no packet loss circumstances, and its effectiveness has been verified by previous researches.

C. Performance Problem in Bandwidth Aggregation on Network Layer

In the case of wireless communication, there are so many packet losses more than the case of wired communication. When Bandwidth Aggregation is operating on Network layer or lower layer, TCP cannot recognize which path causes the packet loss. Thus, TCP executes congestion control and throughput is degraded more than necessary. This is the second problem in Bandwidth Aggregation on Network Layer.

Packet-Pair based Earliest-Delivery-Path-First algorithm for TCP applications (PET) and Buffer Management Policy (BMP) were proposed for the purpose of fixing that problems on Network layer[6].

PET has functions estimating which path should be used more strictly and dynamically. BMP is implemented in receiver node, evaluates whether a received packet is needed to line up or caused packet loss. When BMP receives later sequence number packet, it informs packet loss was occurred for sure. Otherwise BMP delivers correct order packets to TCP.

With PET and BMP, more effective communication is realized compared with implemented EDPF, in particular, when packet losses occur. However, in circumstances with a lot of packet losses occur, even PET-BMP cannot exercise efficient Bandwidth Aggregation. This is most difficult problem to solve in Bandwidth Aggregation on Network Layer. Referenced researches claim it is possible to get expected results with eliminating packet losses using other methods. In reality, it is too difficult to eliminate packet losses in wireless communication.

III. Our Proposal for Bandwidth Aggregation

As shown in previous chapters, we face a various obstacles using Bandwidth Aggregation on Network layer and/or lower layer. Thus, we propose Middleware layer that aggregate bandwidth on the middle of Application layer and Transport layer. Figure 1. shows comparison between Bandwidth Aggregation on Network layer and our proposed model.

A. An Overview of Our Proposal

Our proposal model has some TCP connections per each paths and aggregates their connections. Therefore, applications are not required to be conscious of aggregating bandwidth. It has some TCP congestion windows which prevent throughput degradation more than necessary in circumstances with many packet losses per each paths. Its feature avoid the problem in implementation on Network layer. Our previous research work shows their problems on Network layer are solved[7]. The defect which PET-BMP could not solve is overcome by our method, which means that Bandwidth Aggregation on Middleware Layer is effectively than on the other Layers.

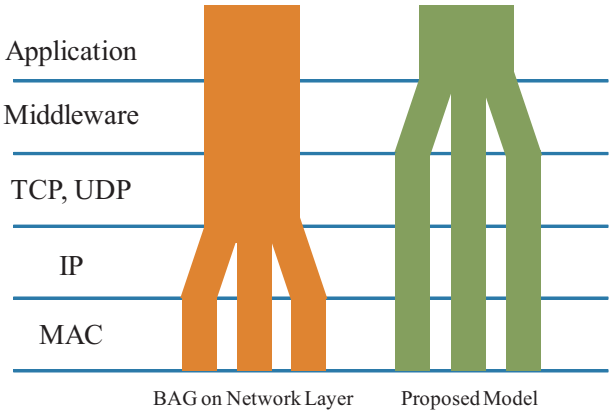


Figure 1. Comparison Between BAG on Network Layer and Our Proposed Model

This approach can also be implemented by modifying TCP which aggregates some connections on Transport layer. However, as an easier way, we can use existing TCP for the purpose of achieving the most efficient Bandwidth Aggregation.

B. The Design of Our Proposed Model

The sender Middleware establishes TCP connections on every possible paths. They receive a packet from an application and give sequence number to a packet. A packet is sent to enabled connection. The receiver Middleware puts received packets in correct order and give them to an appropriate application.

The receiver Middleware has a possibility that some packets arrive by incorrect order and needs to have buffer to restore packets for the purpose of waiting for the packet with expected sequence number. Estimation of required buffer size in each circumstances is one of the important points for designing the Middleware. BMP also considers about buffer size and controls how packets should be deriverd. We propose the method on other layer and suppose that they will behave differently.

IV. Evaluation of Queue Size with Simulation

In this experiments, we are motivated by the advantages that uses Bandwidth Aggregation through simultaneous use of multiple interfaces. We have used simulation software QualNet for their experiments[8].

For the purpose of designing Middleware, the buffer size of Middleware receiver has to be estimated clearly. We have investigated their size under various circumstances.

 ISBN:978-1-61208-113-7

A. Scenario 1 - Low Bit Rate Wireless Communications

Node 1 sends some data to Node2 which has 2 interfaces through 2 paths referring to Figure 2.

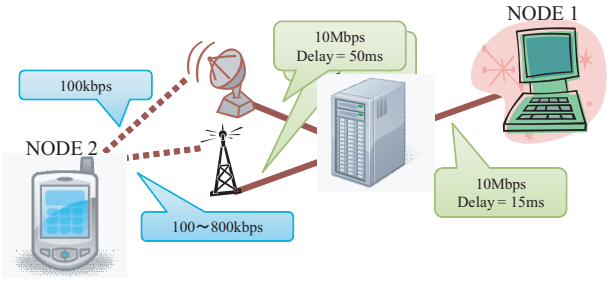


Figure 2. An Overview of Scenario 1

The bandwidths at wired connection are 10Mbps. One of wireless connection is fixed to 100kbps and the other is varied from 100kbps to 800kbps. The rate of two bandwidth of wireless connection is varied from 1:1 to 1:8. Transport protocol is set to TCP new Reno, and parameters are configured by following Table 1.

Table I
TCP PARAMETERS

MSS	1460Bytes
Send buffer	65535Bytes
Receive buffer	65535Bytes

B. Scenario 2 - High Bit Rate Wireless Communications

Node 1 sends some data to Node2 which has 2 interfaces through 2 paths referring to Figure 2.

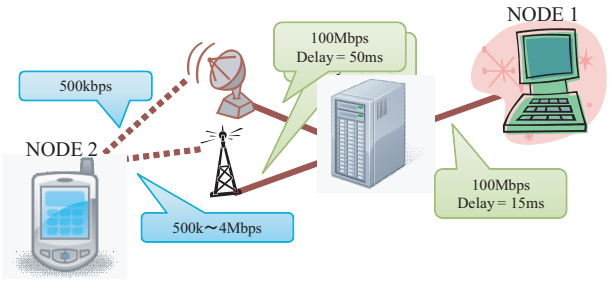


Figure 3. An Overview of Scenario 2

The bandwidths at wired connection are 100Mbps. The one of wireless connections is fixed to 500kbps and the other one is varied from 500kbps to 4Mbps. The rate of two bandwidth of wireless connection is varied from 1:1 to 1:8. Transport protocol and TCP parameters are configured same as Scenario 1.

C. Term of Steady State and Unsteady State

Figure 4. shows throughputs of two connections and buffer size of receiver middleware when bandwidth of wireless connections are set to 100kbps and 300kbps.

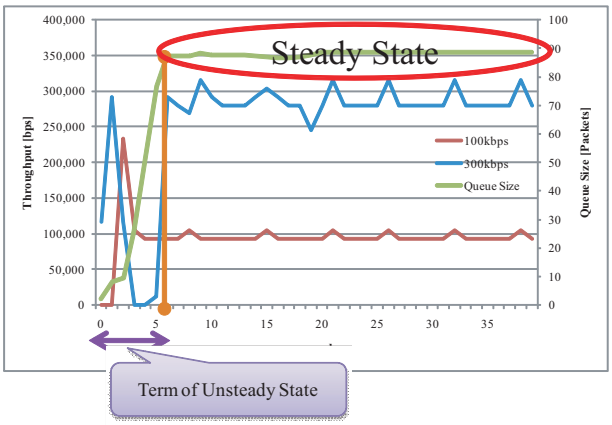


Figure 4. Throughputs and Queue Size

After a short while, two of wireless connection's throughput show efficient communication. Queue size of receiver Middleware is growing at first and settle down at a value. We call the state that buffer size is stable "Steady Term", and the time of until being Steady Term "Term of Unsteady State". We focus on their values at various circumstances.

D. Aasociation Between Rate of Bandwidths and Required Buffer Size

Figure 5. shows buffer size at Steady Term in Scenario 1. when rate of two bandwidths are changed.

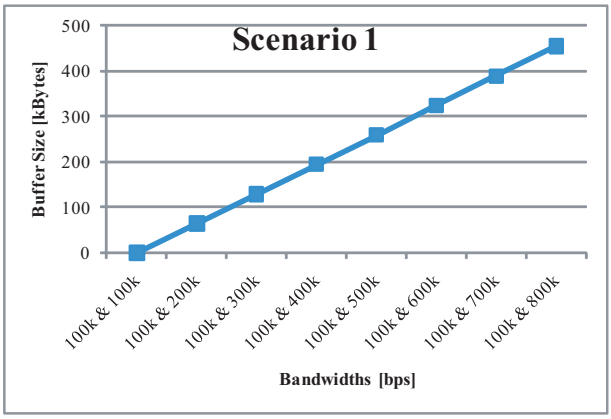


Figure 5. Queue Size in Scenario 1

The value of buffer size when two interfaces has same bandwidths is 0. It is proportional to the ratio of one interface's bandwidth to other interface's bandwidth.

 ISBN:978-1-61208-113-7

Figure 6. shows buffer size at Steady Term in Scenario 2. when rate of two bandwidths are changed.

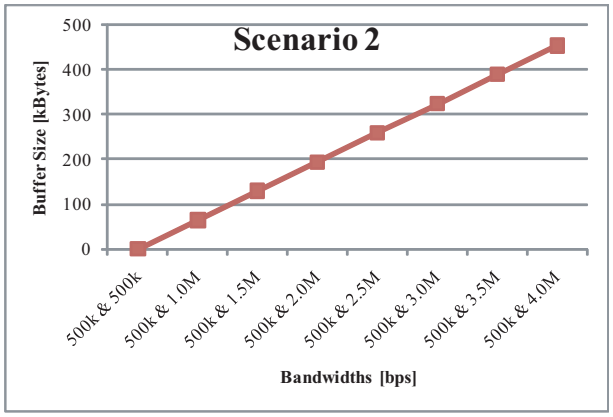


Figure 6. Queue Size in Scenario 2

The value is proportional to the ratio of one interface's bandwidth to other interface's bandwidth as well as Scenario 1. Although Scenario 1. and Scenario 2. have different bandwidth and different rate of bandwidth between wired and wireless, buffer size is resolved by rate of two bandwidth of wireless connection.

E. Association Between Rate of Bandwidths and Time of Unsteady State

Figure 7. shows time of Unsteady State in Scenario 1. and Scenario 2. when rates of two wireless connections' bandwidth are changed.

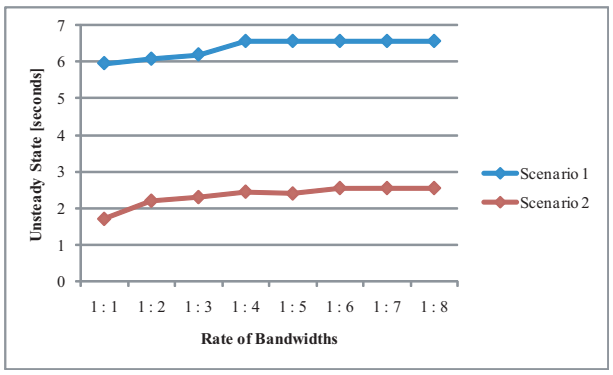


Figure 7. Unsteady Time

The time of Unsteady State in Scenario 2. which has high bit rate is shorter than its in Scenario 1. which has low bit rate in whole. The rates of bandwidth between two wireless connection do not affect their time.

V. Conclusion and Future Work

In this paper, we have experimented with network simulator for the purpose of evaluation of the communication using multiple interfaces simultaneously. The methods of Bandwidth Aggregation on network layer still have problems, for instance, they can not recognize which path cause the packet loss. We have proposed the model of Bandwidth Aggregation on Middleware in order to eliminate their problem. Their effect are verified compared with previous method since we can get comparable throughput as well as aggregating throughput of multiple connection. The receiver Middleware needs to have buffer to restore the order of packets' sequence number. We have investigated how large buffer is needed in various situations. The mobile node which has two interfaces varies one of interface's bandwidth and observes the buffer size. The result shows it proportional to the ratio of one interface's bandwidth to other one.

In the future, we will implement the feature of buffer size that demonstrated by the experiments and function on the sender Middleware considering how to distribute each packets to the paths. Moreover, we will suppose that mobile node can have three or many wireless interfaces and study the result in that cases. In addition, we try to achieve more efficient Bandwidth Aggregation in a various situations, for instance, occurring packet losses, various pattern of lower layer, and dynamically-changed bandwidth.

References

[1] M. Stemm and R. Katz : "Vertical handoffs in wireless overlay networks," Mobile Networks and Applications Vol.3, No4, pp.335-350, Jan. 1998.

[2] IEEE P802.3ad Link Aggregation Task Force : http://grouper.ieee.org/groups/802/3/ad/

[3] Junwen Lai Ming Zhang and et al : "A transportlayer approach for improving end-to-end performance and robustness using redundant paths." USENIX 2004 Annual Technical Conference, pages 99-112, 2004.

[4] H. Nozawa, N. Honda, K. Sakakibara, J. Nakazawa, and H. Tokuda: ARMS: "Application-level Concurrent Multipath Utilization on Reliable Communication" Internet Conference 2008, Oct. 2008.

[5] K. Chebrolu and B. Raman : "Bandwidth Aggregation for Real-Time Applications in Heterogeneous Wireless Networks," IEEE Transactions on Mobile Computing, Vol.5, No4, pp.388-403, April 2006.

[6] K. Chebrolu, B. Raman, and R.R. Rao : "A Network Layer Approach to Enable TCP over Multiple Interfaces," J. Wireless Networks (WINET), Vol.11, No5, pp.637-650, Sept. 2005.

[7] E. Miyazaki, O. Altintas, and M. Oguchi : "A Study of Bandwidth Aggregation Using Multiple Interfaces on Middleware Layer" DICOMO 2010, July 2010.

[8] Scalable Network Technologies : http://www.scalable-networks.com/

 ISBN:978-1-61208-113-7

Mobile Ad-hoc Networks: an Experimentation System and Evaluation of Routing Algorithms

Maciej Foszczynski, Marek Adamczyk, Kamil Musial, Iwona Pozniak-Koszalka, Andrzej Kasprzak
Dept. of Systems and Computer Networks, Wroclaw University of Technology
Wroclaw, Poland
e-mail: iwona.pozniak-.koszalka@pwr.wroc.pl

***Abstract*—The paper concerns the problem of path finding in wireless ad-hoc networks. Several algorithms, including meta-heuristic algorithms, evolutionary algorithm and the created hybrid algorithm, are considered. Algorithms have been implemented into a designed experimentation system. The system allows making simulation experiments along with multistage experiment design. In the paper, the results of some experiments are discussed. Moreover, the comparative analysis of efficiency of algorithms is presented. It may be concluded that the proposed hybrid algorithm seems to be promising.**

Keywords-wireless network; ad-hoc network; path finding; meta-heuristic algorithms; hybrid algorithm, experimentation system, simulation, efficiency

I. INTRODUCTION

Mobile wireless ad-hoc networks are networks with a short period of life. An ad-hoc network is a wireless network to which mobile devices that can act both as client and access point are connected. The most characteristic feature of the ad-hoc network is the lack of any central control device, and also any device to supervise the operation of this information exchange system. Another important feature is the lack of fixed network infrastructure. Systems with this type of connection, therefore, are characterized by high variability and irregularity, which implies the problems absent, or present to a lesser extent in the standard fixed infrastructure networks, both wired, and wireless. Mobility of devices forming such structure is the cause of irregular construction and is a reason of frequent changes in the network structure. The consequence of these characteristics is high importance of algorithms to find not only the shortest path leading from source to destination node, but also to be able to find it fast, regardless of network structure changes. Performance of the algorithm that solves this problem with a large variation of the network structure is crucial, because the algorithm will have to be used after any change in the network structure.

This paper in its content aims to present and formulate the problem (Section II), and demonstrates the variety of its synthetic solutions (Section III). Major emphasis has been made to describe and present the experimentation system created (Section IV), and the results of testing of certain algorithms obtained with this system and using multistage experiment design ideas [1] (Section V). In the final part of the paper, the matter of prospects for the future is raised, including a summary (Section VI).

II. PROBLEM STATEMENT

To fully realize the problem of path finding in a graph of mobile ad-hoc network, one have to imagine a sample network, like the one shown in Fig. 1. It is clear to see, that from a mathematical point of view, this problem can be reduced to find the shortest path between two vertices of an undirected graph.

Figure 1. Sample structure of ad-hoc network.

Mathematical model symbolizing the entire analysed network is a non directed, weighted graph. Vertices in the graph represent individual devices in the network. Connections between the vertices are the physical representation of the wireless connections between devices. The weight of each of the edges in the form of a specific number, defines the quality of the connection. In order to simplify the mathematical analysis of the problem, it can be assumed that the larger the weight, the worse the connection quality. The final element which is necessary to build a full, abstract representation of the problem is to determine the conditions of existence of the connections between vertices.

In the proposed model, the possibility to connect two vertices in the graph is defined by their range, which is an abstract representation of the range of wireless devices in real ad-hoc networks. In the mathematical model, it will also be the number given in standardized units, to determine the radius of coverage of the given vertex. Based on the radius, it can be determined which of the neighboring vertices of a vertex can connect to it and, therefore, can be connected with an edge, what may represent a real connection.

III. THE ALGORITHMS

Two proactive algorithms and two author's reactive algorithms are under consideration, including implementations of Dijkstra and A-star algorithms, as well

 ISBN:978-1-61208-113-7

as ACO (Ant Colony Optimization) and Hybrid algorithms. Dijkstra's and A* algorithms' main purpose was to provide comparison to the reactive algorithm in a modified form of Ant Colony Optimization, and the proposed (by the authors of this paper) hybrid algorithm, which is a combination of modified versions of two of the selected algorithms.

A. Dijkstra and A-star Algorithms

Dijkstra's algorithm is an algorithm that always returns the optimal or close to the optimal route, although it is computationally greedy. In this case, the algorithm has been modified in such way, that after finding the path to the destination node it finishes the path finding process.

Necessary condition for the algorithm is to divide the vertices of a graph into two sets [2]. One set contains the vertices to which paths have been already counted, and the other contains all the nodes which have not yet been processed.

Determination of the path is made iteratively, e.g. [3]. As the first vertex, the initial, start vertex of the simulation is set. In the A* algorithm, like Dijkstra's algorithm, gives the optimal path between two vertices of the graph, but to calculate the path it uses heuristics [4].

The algorithm minimizes the function $f(x) = g(x) + h(x)$ where $g(x)$ is the distance from the start node to the vertex x and $h(x)$ is the path predicted by the heuristic from the vertex x to the destination node. The values of $f(x)$, $g(x)$ and $h(x)$ are stored in three tables [5].

As heuristic functions, we have chosen the „Euclid" function (1), and „Manhattan" function (2).

$$h(x) = \sqrt{(x.X - end.X)^2 + (x.Y - end.Y)^2} \quad (1)$$

$$h(x) = |x.X - end.X| + |x.Y - end.Y| \quad (2)$$

Determination of the path is iterative, as in Dijkstra's algorithm e.g. [6].

B. Ant Colony Optimization Algorithm

The idea of the ant colony optimization is to base the algorithm's work on the behaviour of the colony of ants, seeking a route from their nest to food source and back again, e.g. [7].

Ants, as they move along the edges of the graph, leave their pheromone to indicate to the other ants that the edge has already been visited [8]. With time, the concentration of pheromone P_c on the edges of the graph is decreasing with concentration loss factor l, i.e. *new* $P_c = P_c \cdot l$.

Pheromone concentration loss process is continuous and occurs at the beginning of each run of the algorithm's iteration, e.g. [9].

The proposed modification of a classic ACO consists in dividing ants into two categories: forward and backward ants. Forward ants' main purpose is to explore the graph and to find the destination node. When forward ant reaches the destination, it sends back backward ant and dyes. Backward ants are much more likely to follow the pheromone, because their priority is to consolidate the route and get back to the source node quickly, from where they send forward ants again.

In a classic implementation of this algorithm, routing tables are used to locally memorize the results of the algorithm's work in the network. For the means of an abstract implementation, routing tables have been omitted, as assumed that the subject of the research was the path finding itself, rather than maintaining the route within a given instance of the problem.

Determination of path length in this algorithm is made in an iterative manner. The path which ant chooses for the next step is added to the total value for each ant. Final result is determined as the shortest path of all of the ants.

C. Hybrid Algorithm

Hybrid algorithm is an author's algorithm, which was developed in response to the need to reduce the cost of finding the path, regarding the implementation of the first n steps as quickly as possible, and then, after a quick advancement in path selection in the first stage, further optimization of the path made by using one of specialized algorithms e.g. [10].

To implement this algorithm, modified version of ACO was implemented in conjunction with Dijkstra's algorithm e.g. [11]. Modification has been made to limit the amount of ants and to modify the way the ant chooses its next vertex in the graph. Algorithm obtained in this way allows for a close to random, but relatively controlled first n steps, which will be made. After completing n steps, the ACO finishes and passes its current vertex as the starting vertex for the next algorithm.

After the calculation of the initial direction, Dijkstra's algorithm is run, which is aimed to find the path to the destination node if it has not been reached yet e.g. [12].

Path length in this algorithm is made in an iterative manner, as a sum of path values given by both of the algorithms.

IV. EXPERIMENTATION SYSTEM

A. Basic Characteristic

The Windows platform has been chosen as an implementation environment, on which an application in C# programming language has been created. To run the simulator, the workstation must be equipped with Windows 2000/XP/Vista/7 operating system and .NET Framework 3.5.

The simulator has an interface that allows the user to easily configure all the parameters of the application. Moreover, its construction allows to quickly and easily extending its capabilities, including possible addition of new algorithms.

B. Functiona Features of Application

After launching the simulator application, the application main window appears, as shown in Fig. 2. The main window is divided into clearly separated areas.

 ISBN:978-1-61208-113-7

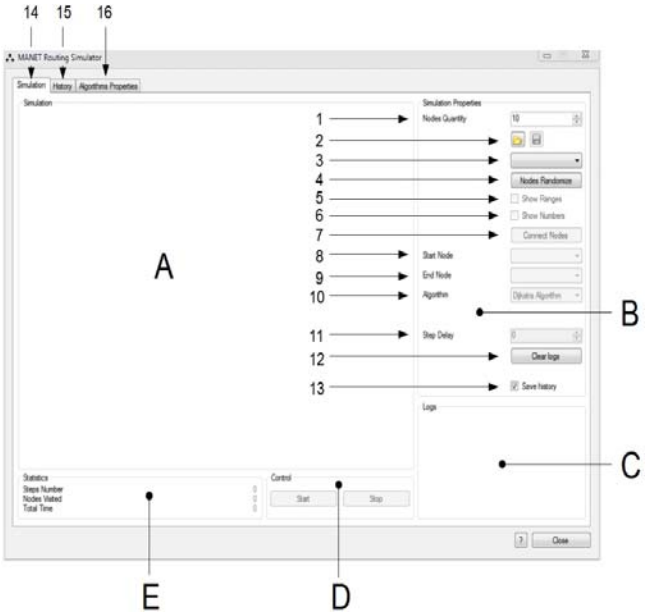


Figure 2. Main window of application

The largest area of the application window is the area of simulation (A). In this area the graph representing the specific problem and the effect of the algorithm will be shown. It is also possible to modify a specific instance of the problem before running the algorithm itself.

In the settings area, (B), we see the basic parameters that can be modified in the program. The first one is the parameter determining the number of vertices in the graph, which is to be generated (‘1’)

Next are two buttons, allowing to save the current graph to a file and load a saved graph to the program (‘2’).

A select form (‘3’) allows choosing a specific instance of the problem saved earlier. To add a graph to the list, save it in a subdirectory called „Graphs” in the root directory of the simulator. After adding the file and restarting the application, saved graph appears on the defined graphs selection list.

Under the selection of defined graphs, we see the graph draw button (‘4’), allowing to generate a random graph, consisting of the number of vertices determined by the parameter (‘1’).

Vertex positions are random according to normal distribution. If the arrangement of the vertices is not satisfying, it is possible to draw another instance by re-clicking on the „Randomize” button, or manually modifying the position of given nodes. Nodes in the simulation area can be moved using drag-and-drop method.

Below are two fields that allow interfering in the amount of information displayed in the simulation. „Show ranges” select (‘5’), displays the circle around each of the nodes, symbolizing node's range in relation to the other vertices. Selecting „Show numbers” parameter (‘6’) will cause a number to appear next to each node which enables its identification.

The number (‘7’) in the illustration has been assigned to a button that connects all vertices in the graph. Connections are made on the basis of nodes range. The connection between the two vertices *a* and *b* may occur if, and only if, the range *r* of the vertex with less value is less than or equal to the distance d_{ab} between the vertices (3).

$$C(a,b) = \begin{cases} 1 : \max(r_a, r_b) \geq d_{ab} \\ 0 : if\ else \end{cases} \quad (3)$$

The next two fields, (‘8’) and (‘9’), allow the selection of the source and destination node in the graph. Algorithms will find the shortest path between the initial and final vertex, using only the available connections. There is a possibility that it will be impossible to find any path between two selected vertices.

After selecting the initial and final vertex, an algorithm that will look for the shortest path between them can be chosen. Selection of the algorithm takes place by selecting from the drop-down list (‘10’).

If the algorithm supports additional parameters for its operation, before the start of the simulation it is possible to configure the parameters in „Algorithm Properties” (‘16’).

The last parameter that can be set is the „Step delay” (‘11’). Here the number of milliseconds that the simulator will wait after each step of the algorithm can be specified. Note that due to the large variety of algorithms, this parameter is purely indicative.

Additional button „Clear logs” (‘12’), is used to delete the exported results of the algorithm run.

The last option available in the main settings area is a field which allows enabling or disabling algorithm run history (‘13’). When this option is enabled, step-by-step algorithm history analyse is possible in the „History” tab (‘15’).

Algorithm results field (C) is located under the main settings area. Basic results of algorithm run are shown in this field.

Below the simulation area two buttons marked „Start” and „Stop” are located (D). These buttons allow starting and stopping the simulation.

Current algorithm run information is shown in the live statistics area (E). These statistics are updated with every step of the algorithm, so if the delay of the algorithm iteration was set, it will be possible to analyse statistics during the run of the algorithm.

C. Concept of Research

Implementation environment allows for testing of the algorithms in several aspects. The index of performance treated as the measure of the efficiency, is the overall quality of the path d_i, which is obtained as a result of the algorithm run. The target function is expressed by (4).

$$F_c = \sum_i d_i \quad (4)$$

 ISBN:978-1-61208-113-7

At the same time, the algorithm should visit the least amount of vertices possible, and take the smallest amount of time for its action. Number of vertices visited by the algorithm and the time of the execution are associated with its actual demand for resources and traffic generated by the algorithms in the network, therefore the quality of these parameters is not left without a meaning to the estimation of the quality of functioning of the algorithms.

Remaining at the level of abstract simulation of the behaviour of algorithms for searching paths in the graph, the quality of paths and quantity of visited vertices is taken into account and in this respect, the algorithms are compared.

V. INVESTIGATIONS

A. Research Theses

It is estimated that Dijkstra's algorithm provides an optimal, or very close to the optimal solution, but obtains it at great expense of calculation, which should result in relatively long run time. In the real network environment, the additional disadvantage of this algorithm is the need to process the entire graph each time a request to find the appropriate path is sent.

A* algorithm, based on the heuristic methodology, as a result of its action finds the optimal solution to the problem, using relatively large amount of resources to obtain it, so it predictably is to visit a large number of nodes in the graph.

Another approach to the problem is presented by the Ant Colony Optimization which in contrast to the other algorithms can run in the network for a long period of time, gradually improving the result and adapting to various network structure changes. In its abstract implementation, this algorithm should not show up in finding the optimal path, since the run time has been limited. Noteworthy, in the real implementation of the algorithm it exhibits a high degree of flexibility to adapt to rapidly changing network topology.

Experimental implementation of the hybrid algorithm is an interesting subject of research. It is difficult to accurately predict the algorithm behaviour and possible results, but according to the assumptions, the algorithm is to provide relatively satisfactory outcome in the short period of time, while showing a small number of visited vertices.

B. Experiment Design

Each algorithm was tested for five different total numbers of vertices in the graph. Instances of graphs with 20, 30, 50, 70 and 100 vertices were chosen, and saved in order to provide the same test environment for each of the algorithms. For each of the numbers of vertices in the graph and the values of parameters of each algorithm, 10 measurements were made, what allows to objectively asset the quality of the results, thus calculating the average results for each of the algorithms.

The experiment design, constructed along with the multistage experiments concept [13], was composed of the series of series of single executions of algorithms. The detailed values of the flexible parameters are specified in Table 2. It is necessary to mention, that all experiments were conducted in the environment described in the previous subsections.

TABLE 2. Experiment Design.

Algorithm	Parameter	Number of vertices				
Dijkstra	-	20	30	50	70	100
A*	Euclid	20	30	50	70	100
A*	Manhattan	20	30	50	70	100
ACO	P_c = 0,0004	20	30	50	70	100
ACO	P_c = 0,0016	20	30	50	70	100
ACO	P_c = 0,0064	20	30	50	70	100
ACO	P_c = 0,0128	20	30	50	70	100
Hybrid	n = 5	20	30	50	70	100
Hybrid	n = 10	20	30	50	70	100
Hybrid	n = 20	20	30	50	70	100

C. Results and Discussion

In the first case, the thesis, concerning the efficiency of Dijkstra's algorithm, was taken under consideration. Performed simulations of the algorithm run time for 100 vertices, shown in Fig. 3, confirm the assumption that the algorithm is characterized by a relatively low efficiency, needing a lot of time to process all the data.

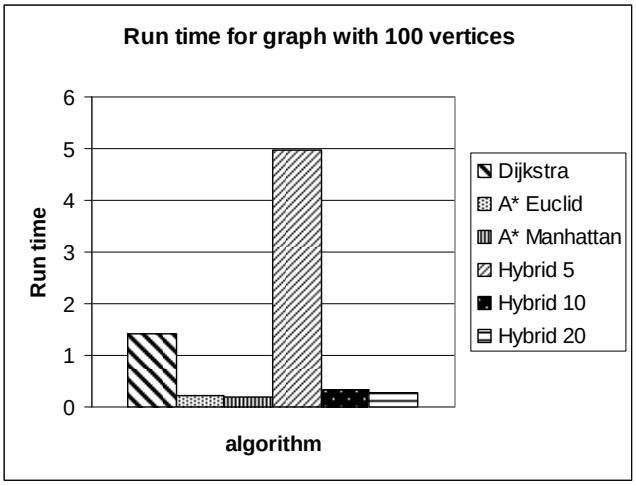


Figure 3. Run time of algorithms for 100 vertices.

It is worth to be mentioned, that a high processing time has also been obtained for the hybrid algorithm, which greater part for the graph of 100 vertices is Dijkstra's algorithm, which further confirms the truth of stated thesis.

A* search algorithm, due to the complex structure of the implementation using the heuristic methods, has proved to visit the largest number of vertices, which confirms the related thesis. Example of the number of visited nodes for the graph of 30 vertices, shown in Fig. 4, classifies it right after the Ant Colony Optimization, which in the actual

 ISBN:978-1-61208-113-7

implementation is intended to work without the time limitation.

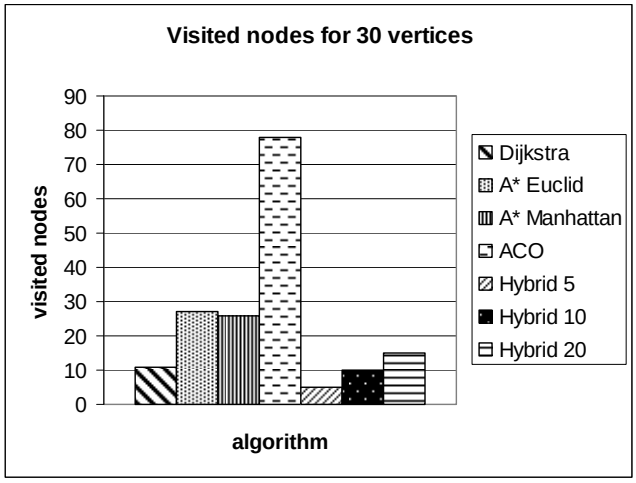


Figure 4. The total number of visited nodes for 30 vertices.

A noteworthy fact is that irrespective of the type of used heuristic function, A* algorithm, according to the thesis, is characterized by a large number of visited vertices, and so, in fact, a large number of generated connections, but generating the optimal solution of the path finding problem.

According to the thesis set for the Ant Colony Optimization, it did not provide optimal results, however, it is able to adapt to the network structure. Fig. 5 shows how the path quality obtained by the ACO differs from the quality of paths developed by other algorithms in adequate run time. Clearly, author's ACO algorithm is able to find very good quality path and is further characterized by very high flexibility of action.

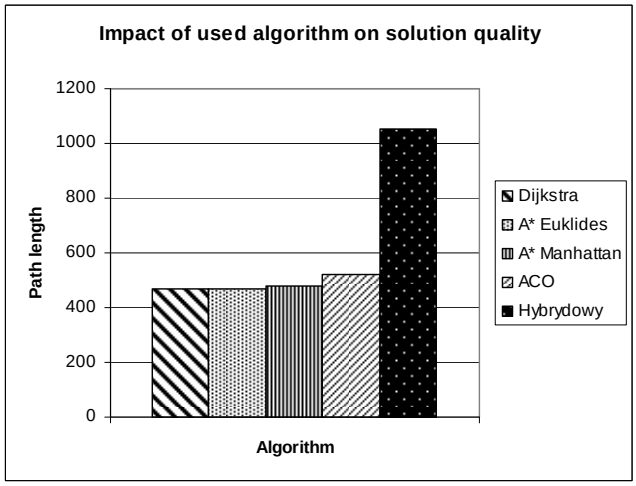


Figure 5. Path length for 30 vertices.

It is worth to note, that the quality of path obtained by the ACO changes with the pheromone concentration loss factor. Fig. 6 shows, that properly chosen pheromone loss factor can help to make the algorithm even more effective.

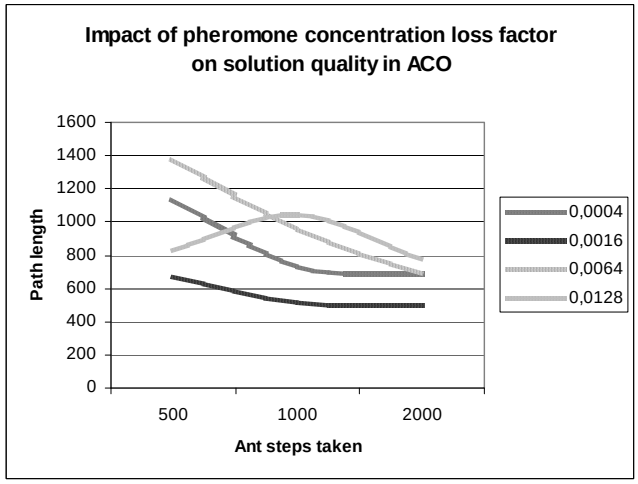


Figure 6. Impact of pheromone loss on solution quality in ACO.

The results of an experimental hybrid algorithm proved to be a confirmation of assumptions of its possible behaviour. With the increase in the contribution of modified Ant Colony Optimization, which means increasing the importance of the pseudo-random part of the algorithm, hybrid algorithm significantly increased the speed of its operation.

As shown in Fig. 7, the implementation of the first 10 steps using the modified ACO resulted in a drastic reduction of the algorithm run time, at the cost of decreasing the quality of the solution.

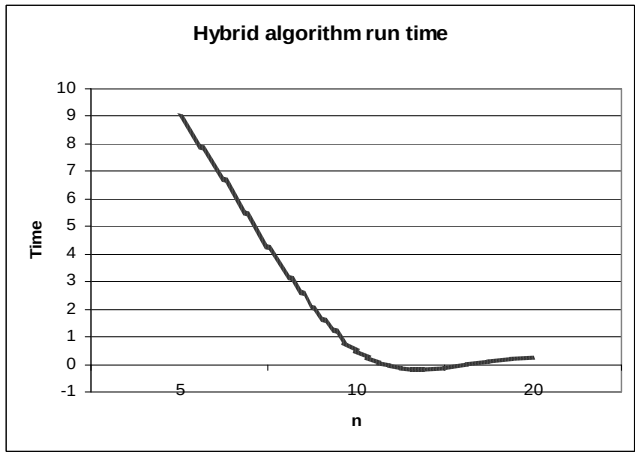


Figure 7. Hybrid algorithms run time.

With the increase of the *n* parameter, the number of steps taken by the algorithm has significantly decreased. The dependence is shown in Fig. 8. Number of visited vertices remained more or less stable, which further emphasizes the importance of pseudo-random part of the algorithm to reduce the amount of the calculation.

 ISBN:978-1-61208-113-7

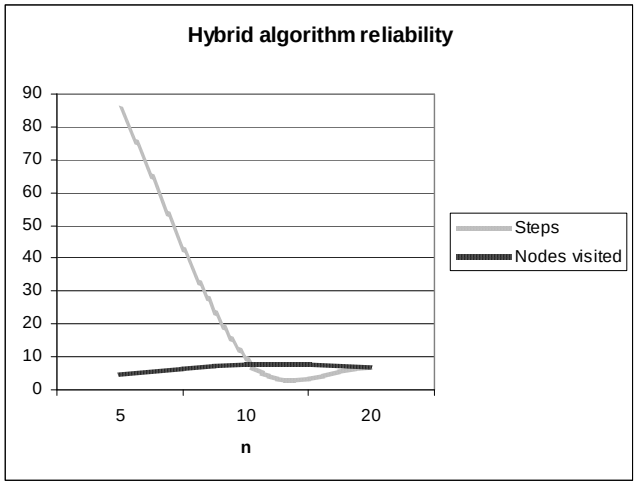


Figure 8. Hybrid algorithm reliability.

Close to random nature of the hybrid algorithm is stressed by the fact that for the *n* parameter value equal to 20, the number of performed steps has slightly increased, which is caused by too much involvement of the random part of the algorithm. Appropriately balanced algorithm parameters can improve the overall quality of obtained results and the algorithm itself provides promising results and a solid basis for further research and development.

VI. CONCLUSIONS

Research carried out allowed drawing far-reaching proposals for the design of systems based on the idea of finding a path in wireless ad-hoc networks.

Diversity of the algorithms realizing the routing in wireless ad-hoc networks available to implement requires clarifying and clearly specifying the system requirements. When it is known that the system must be resistant to changes in network and rapid adaptations to new conditions, it is advised to use algorithms that provide the desired flexibility, for example, Ant Colony Optimization algorithm. If the key is to obtain a satisfactory solution to the problem in the shortest time possible and subjects minimize the consumption of resources, a good solution could be a hybrid algorithm, similar to the algorithm proposed in this paper, which can combine the best features from selected algorithms while maintaining an appropriate balance between their drawbacks.

In the future implementation of similar project, the right direction would be to develop the idea for giving possibilities of simulations closer to the reality, gradually to move away from abstract approaches. This would enable more specific implementation of the algorithms for selected problems and to conduct more in-depth research. Nodes could use the parameters of the actual nodes of ad-hoc network, which combined with assigning more details to the connection between two nodes would increase the level of realism, which would help to carry out further tests, developing more accurate reflection of reality.

The computer experimentation system presented in this paper was designed with a possibility to expand it with additional modules. Increasing the functionality and reducing the level of abstraction can provide a solid basis for future research in this topic.

REFERENCES

[1] L. Koszalka, D. Lisowski and I. Pozniak-Koszalka, “Comparison of Allocation Algorithms with Multistage Experiments”, Lecture Notes in Computer Science, vol. 3984, Springer, 2006, pp. 58-67

[2] E. W. Dijkstra, “A Note on Two Problems in Connexion with Graphs”, Numerische Mathematik, 1959.

[3] A. Kasprzak, “Packet Switching Wide Area Networks”, WPWR, Wroclaw, 1997 /in Polish/.

[4] M. Abolhasan, T. Wysocki, and E. Dutkiewicz, “A review of routing protocols for mobile ad hoc networks”, University of Wollongong, 2003.

[5] N. Wirth, “ Algorithms + Data Structures = Programs”, Prentice Hall, 1976.

[6] M. K. Marina and S. R. Das, “On-Demand Multipath Distance Vector Routing in Ad Hoc Networks”, University of Cincinnati, 2001.

[7] M. Dorigo and T. Stützle, "Ant Colony Optimization”, MIT Press, 1997.

[8] C. Blum, “Ant colony optimization: Introduction and recent trends”, Physics of Life Reviews, 2005.

[9] M. Dorigo, “ *Ant Colony Optimization* ”, Scholarpedia, 2007.

[10] Z. Michalewicz, “Genetic Algorithms + Data Structures = Evolution Programs”, Springer, 1996.

[11] A. Botea, M. Muller, and J. Schaeffer, “Near Optimal Hierarchical Path-Finding”, Journal of Game Development, 2004.

[12] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, “Dijkstra's algorithm”, Introduction to Algorithms, Section 24.3, MIT Press, 1990.

[13] D. Ohia, L. Koszalka, and A. Kasprzak, “Evolutionary Algorithm for Congestion Problem in Computer Networks”, Springer, Lecture Notes in Artificial Intelligence, vol. 5711, 2009, pp. 113-122.

 ISBN:978-1-61208-113-7

Home Automation with IQRF Wireless Communication Platform: A Case Study

Vladimir Sulc
MICRORISC s.r.o.
Jicin, Czech Republic
sulc@microrisc.com

Radek Kuchta, Jaroslav Kadlec
Faculty of Electrical Engineering and Communication
Brno University of Technology
Brno, Czech Republic
kuchtar | kadlecja@feec.vutbr.cz

***Abstract*— This paper describes a new wireless communication platform IQRF. In the paper is description of main features of the platform, available communication modules, gateways to other wired and wireless communication system, and development tools. The paper also describes a case study focused on IQRF Smart house concept. IQRF platform was designed and developed especially for home automation and telemetry projects.**

Keywords- Home Automation; IQRF; IQMESH; Wireless communication; Networking.

I. INTRODUCTION

Smart House is a building, equipped with electronic system enabling occupants to control, program or use variety of electronic devices by entering a simple command. For example, a homeowner on vacation can remotely monitor, arm or disarm security system or switch on heating system when he is returning from his vacation earlier. Devices can also communicate to each other. Remote thermometer located in the place of comfort can provide data to Heating Ventilating Air Conditioning (HVAC) system in different rooms and actuators can interpret variety of different commands sent by Control Unit or by other devices in distributed control systems.

Such electronic systems are making buildings smarter. They consist usually from electronic devices providing data (sensors), devices interpreting control data (actuators), control devices (central units) and from devices providing communication interface to the system (gateways). All of these devices are usually located at different places of the building, therefore there is a need to enable simple, usually low data rate, communication between devices.

In new buildings connectivity would be easily realized by structured wiring during building process, on the other hand in existing buildings it would be a problem to make new wirings, especially when the building has not been prepared for that. Also, from the point of the installation costs additional installation of structured wiring in existing buildings is highly expensive. Cost of wire installation for simple electronic devices, like light switches, is 20 or 30 times higher than the cost of the switch. In this case wireless communication would be an ideal solution. Prices of Radio Frequency (RF) modules are dramatically falling down, enabling widely penetrate the market. The main idea of a house with remotely controlled equipment is shown in Fig. 1.

More complex or larger buildings bring new challenges for wireless communication to cover all devices by the signal which guarantee sufficient QoS (Quality of Service) parameters. Simpler wireless network topologies such as star would be efficiently used for smaller buildings; increase of the transmit power and/or sensitivity of the receivers would help to cover less accessible places, but this approach increases also RF radiation, interferences and it would not be even possible in some cases to cover the whole building due to the obstacles or walls construction. Higher bands (2.4GHz or higher) would face such problems with signal propagation more often than sub GHz bands. Wireless Mesh Network therefore seems to be an ideal communication topology to make buildings smarter.

There are available different wireless communication solutions from different vendors on the market place. These solutions support different network topologies. Many of them are based on 802.15.4 [1] standard defining Physical Layer (PHY) and Media Access Layer (MAC) for Low Rate Wireless Personal Area Networks (LR-WPAN). In most cases they work on non-licensed wireless communication bands. Non-licensed bands are different in a lot of countries. In European Union, there are 433 MHz, 868 MHz, 2.5 GHz and other bands. In the United States of America, there are especially 916 MHz and several others.

One such standardized protocol that works on non-licensed bands is, for example, Zigbee. It involves a solution based on the IEEE 802.15.4 standard [1] prepared by Zigbee Alliance [2]. This standard was developed by consortium of industrial companies especially for building automation [3,4]. There are also special applications for industrial control [5,6,7,8,9,10,11]. Among the proprietary solutions, reference can be made to the technology of MiWi launched by Microchip Technology Inc. [12]. MiWi is based on the aforementioned standard but simpler than Zigbee from the implementation point of view. This technology does not support direct cooperation with Zigbee devices [13,14]. From other solutions available on the market, mention would be made, for example, of the solution promoted by Z-wave alliance [15,16].

These solutions have disadvantage in attempt on being a universal solution targeting every kind of applications. It brings heavier protocols, more difficult and more expensive implementations.

Implementation of solutions such as Zigbee or MiWi consists of software solution stack and hardware solution used for communication. Software solution stack is developed by a microcontroller manufacturer for defined microcontroller or by a producer that wants to supply his products for communication modules designed for the area of domestic automation. The software stack is a package of

 ISBN:978-1-61208-113-7

program routines, functional components and program subsystems (hereinafter Stack) permitting the basic operation of the communication module according to the chosen solution for wireless communication. The manufacturer of the end device uses the modules for selected communication solution, and then creates a further application extension to implement the actual application functionality of the end device [17,18,19,20].

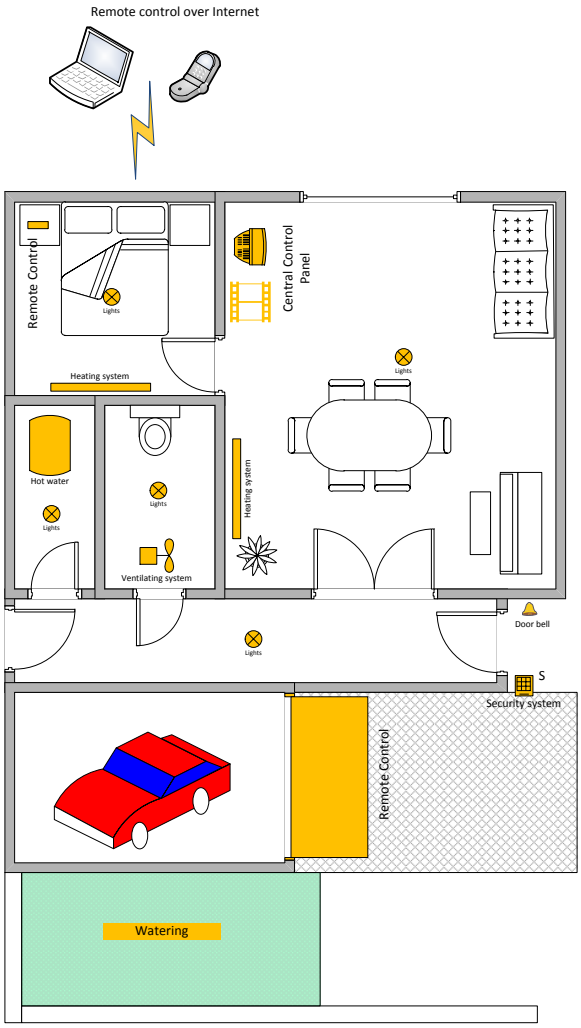


Fig. 1. The main idea of Smart House

There are also other proprietary solutions with wireless functionality. In the most cases they allow simple direct communication link without any other network functionality and they usually use master/slave communication model.

The paper is focused on description of IQRF wireless communication platform and its possible application in the smart house concept. At the beginning of the paper is short description of the smart house concept. Following section describes wireless communication platform IQRF, communication modules, IQRF operating system that allows rapid application development, available gateways to other wireless and wired systems, and description of development tools for fast design and development. At the end of the paper future work and conclusion are summarized.

II. IQRF SMART HOUSE CONCEPT

The main idea of IQRF Smart House concept is shown in Fig. 1. The idea is to allow control almost all electrical equipment through wireless communication. The main parts are lights, heating system, ventilating system and hot water systems.

Smart house concept also contains security system that controls the main entrance doors, windows and when it is needed also contains motion sensors.

The concept also contains global garden watering system and control of garage door and car entrance.

III. WIRELESS COMMUNICATION PLATFORM IQRF

To address requirements from home automation and telemetry systems a new wireless communication platform IQRF was designed. The name IQRF is an acronym Intelligent Radio Frequency. At the beginning the platform was used especially to control electrical heating systems in a hotel or other commercial buildings where centralized control is needed. Now IQRF is designed to control whole set of devices used in a home automation process. The platform was developed by Microrisc company [21]. The main parts of the platform are covered by Czech and US patents [22,23,24,25]. These patents cover a method of creating a generic network communication platform, special signal coding scheme, and direct peripheral addressing in wireless network.

IQRF is using its own concept of the communication module structure. Wireless part is based on short-range radio components produced by RFM Company, which work in non-licensed communication bands. IQRF communication modules are available for 868 MHz and 916 MHz frequencies.

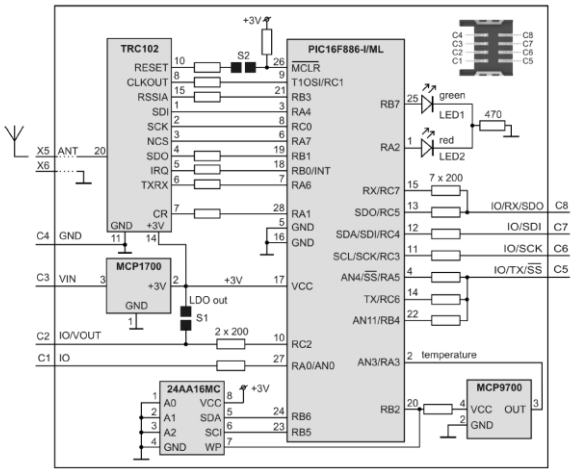


Fig. 2. The block structure of the IQRF communication module

A. Transceiver Module

Whole platform is based on transceivers modules. Basic block structure of a module is shown in Fig. 2.

Transceiver module is a tiny intelligent electronic board with complete circuitry needed for realization of wireless RF connectivity. Microcontroller with an inbuilt operating system, providing debug functionality, integrated LDO regulator and temperature sensor dramatically reduce time of application development. Low power consumption predetermines these modules for use in battery powered applications.

Depends on module version different microcontrollers are used. The newest version is using microcontroller Microchip PIC 16F886. Modules without integrated antenna are the same size like SIM card and they are using the same connector. Modules with integrated antenna are bigger by antenna but still with the same connector for assembling/plugging these modules to a superior systems. Therefore it is possible replacing each other according to the application needs.

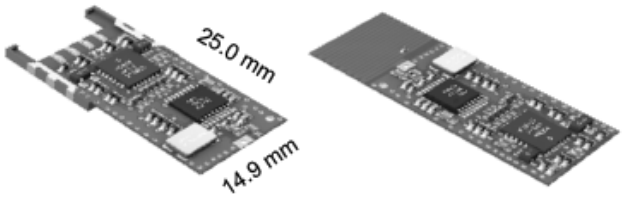


Fig. 3. IQRF communication modules with and without integrated antenna

Basically, the IQRF communication module has three standard input/output interfaces, one analogue input, an SPI interface, and digital ports. Each module contains integrated analogue temperature sensor, LED and 3 V linear regulators, which supplies communication module and moreover can be used for supplying user application. In Fig. 3 is shown the latest version of the IQRF communication modules TR-52 with and without integrated antenna. These modules are using FSK modulation and they have more digital input/outputs.

B. Operating system

Every IQRF Transceiver module is equipped by operating system (IQRF OS) implementing its basic functionality. IQRF OS is buffer oriented. Block scheme is shown in Fig. 4.

IQRF OS dramatically simplifies design phase, programmer of application can focus on application only, detailed study of RFIC and data processing before TX or after RX is not needed. Besides basic functionality IQRF OS provides also mechanism for application upload when the application is compiled. Programmer will set IQRF module to the programming mode, then, via SPI interface application code is uploaded to the module.

Whole system offers about 40 functions. A function block diagram is shown in Fig. 4. The main functions of OS are:

- RF functions for transmitting, receiving, bonding and setting up,
- IIC and SPI communication functions,
- EEPROM access functions,
- three buffers for RF, COM and INFO are available,
- other auxiliary functions for LED, OS information, delays and sleep mode functions are available too.

Up to 64 bytes is possible to send in one packet. The packet size is variable and should be set before packet is sent by a transmit function.

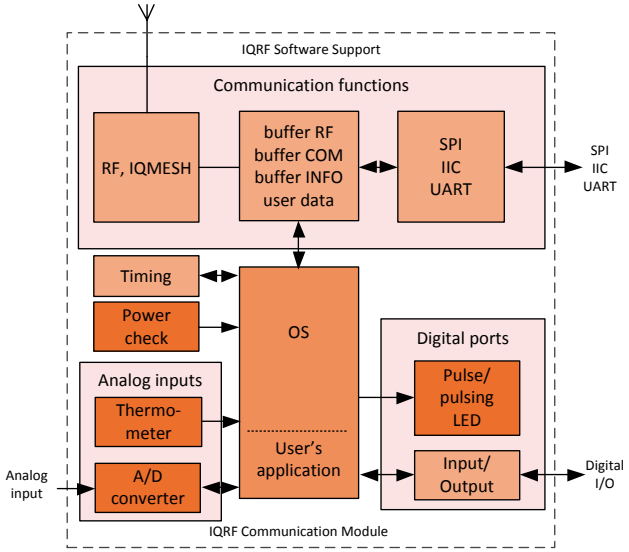


Fig. 4. Basic functionality block diagram of IQRF Operating system

IQRF operating system is implemented to the program memory of the microcontroller. Program memory is divided to two main parts. The first part is used by IQRF operating system and the second is available for user's application. When user's application needs to call some OS function, it calls function address defined in the definition file of the selected OS version. Programmers of the application can use whole set of the microcontroller instruction. Some restrictions for direct program memory access are applied. Because direct program memory access instructions are not allowed in the user's code, IQRF has implemented functions to store and read data from the on chip integrated EEPROM memory.

IQRF is wireless communication platform, therefore IQRF OS supports functions to create network, with different topology. When IQRF networking functionality is used, there have to be one coordinator in the network. Each communication module can work as a coordinator. Each module is possible to connect to two different networks. This functionality allows interconnection of the separated network without any gateway. Network possibilities and MASH functions are described in detail in the next chapter.

To support wireless and network functionality three data buffers are available. The OS also offers functions to copy data between buffers. Buffer called RF contains wirelessly received data or data to be transmitted. COM buffer is used to send and receive data via SPI, IIC and UART interface. INFO buffer is used by system for block operations.

A special signal coding scheme brings higher data throughput due to real time data compression and also higher

 ISBN:978-1-61208-113-7

reliability and noise immunity due to perfect DC balance of the coded signal [24].

OS also offers functions for timing, power control, reset and integrated LED control. Detailed description of all IQRF OS function is in [21].

C. *Gateways and Development tools*

Various gateways to common standards such as Bluetooth, ZigBee and GSM are available. Simple applications can use RS-232 gateway or more useful USB gateway. These simple gateways were developed to allow connection between IQRF and other proprietary solutions. They also allow connecting IQRF and standard PC with user's application.

For more sophisticated applications, GSM or Ethernet gateways are available. To allow interconnection between IQRF and standard wireless solution a Bluetooth and ZigBee gateways are available.

Development tools allow debugging and testing of user applications using supporting software. To provide comfortable environment for a transceiver development kits typically contain interface connectors, battery, interface to user pins and so on.

There is also integrated development environment IQRF IDE that is available for all IQRF development kits. This IDE allows software development with integrated BKND compiler, programming of all IQRF modules. The IDE integrates user application debugging information, SPI communication debugging.

IV. A CASE STUDY

In the next sections, two use cases are described. The first one describes a smart house in vocation program. The second one describes a situation, when owner has to return from vacation earlier.

For our use cases we prepared a small smart house with only a few remotely controlled devices. The central control unit is, in this case, Smart House Central Server with USB – IQRF gateway. This server is possible to program to control whole house. It also allows automatic processes like heating system control or switch on/off lights in selected time periods. The server works also as a network coordinator.

There is selected number of lights. Each of them has IQRF transceiver and is controller remotely through IQRF network. Each light has own remote switch, but in our scenario these switches are not important.

The house is equipped with wirelessly controlled heaters and boiler for hot water.

Because some devices are out of range of Central Server we have to use IQRF Router. Heater 1 is out of range from central server and from the router, but is in range of Light 1. Because each IQRF Transceiver can work on background as a router, we do not need another router.

Whole system is controlled and programed through IQRF Remote Control. This device has graphical LCD and allows direct control of each unit or programming of Smart House Central Server.

In use cases is also a user. The user has access to IQRF Remote Control and has cellular phone with IQRF GSM Gateway phone number. He also knows how to control and program whole house remotely.

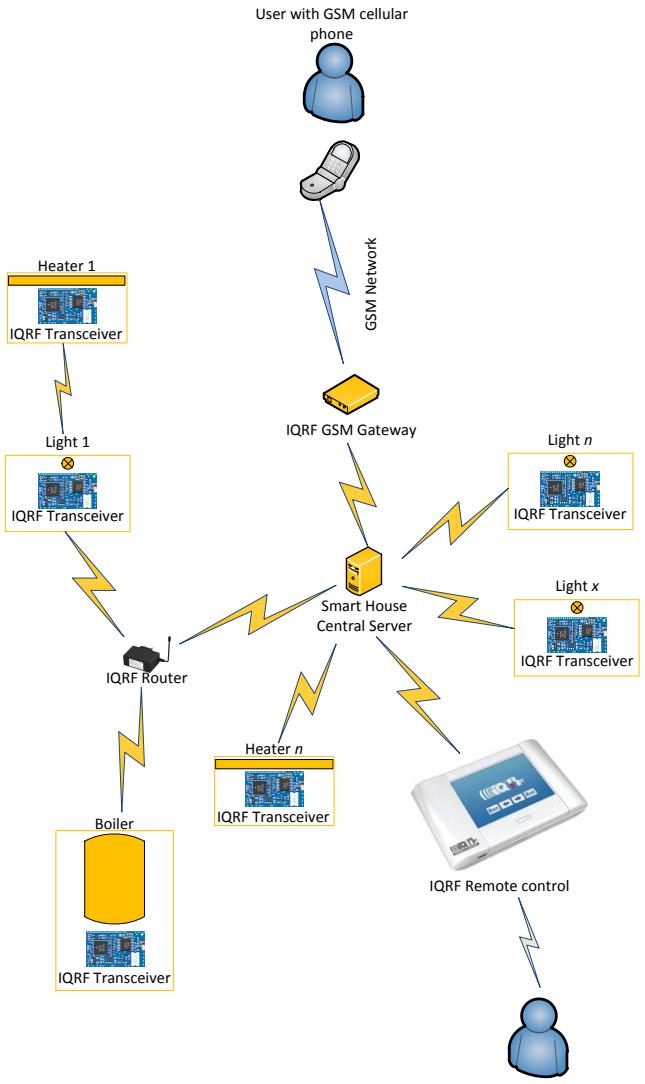


Fig. 5. Block diagram used for a case study

The house also has IQRF GSM Gateway that allows connection from GSM network. The web server is not included in our scenario, but it is another way, how to control whole system.

Network components and their connections of our smart house are shown in Fig 5. Whole communication ways are setup and system is in standard state.

A. *Use Case I*

It is winter time and the user is going out for vacation. Because nobody will stay in the house for vacation time, it is not necessary to have hot water; also temperature in the house should be lower than in standard time. User will create a new profile called "vacation" through IQRF Remote Control. In this profile will select that boiler is switched off

and heating system is switched to tempering only. This profile will store to Central server for future use.

Because on vacation time the house is empty, it is a good time for rubbers. Due to this reason the user will extend vacation profile with automatic light switching. He will select time of switching on and time of switching off for each light. There is also possibility to add random variable for each light. So time will be little bit different each day.

Before the user leaves the house, he will select vacation profile on the remote controller. He also selects the day of return when profile will be automatically changed to standard profile. It is important value, because the heating system has to change temperature from tempering to standard temperature and it takes time.

If the user forgets to change profile he will use a cellular phone to change the profile remotely.

When the user comes back home from vacation he has the same comfort like he leaves home, but during his vacation the smart house saved energy, costs and of course living environment.

B. Use Case II

This use case expands use case I. The smart house is using profile vacation and user is somewhere out. But he has to change his plans, and return back to home earlier.

In this case, he will use his cellular phone to connect through IQRF GSM Gateway to Smart House Central Server. By this way he will change profile from vacation to standard. Internal house system will turn on hot water system and also heating system will try to reach standard temperature immediately.

V. FUTURE WORK AND CONCLUSIONS

IQRF is a new wireless communication platform especially designed and developed for specific requirements from home automation and telemetry. One of the main aims was to offer wireless platform to developers of the end user devices that allows rapid development without necessity of stack implementations.

One of the typical application usages of IQRF is in Smart houses and similar projects. The platform was designed especially for home automation and telemetry projects. Network functionality, available gateways and easy implementation to user devices allow rapid application development without long study period of chosen wireless solution. Developers only use prepared OS functions and work with application layer of communication protocol.

Now we are working on implementation of all features of smart house concept. Patented direct peripheral addressing in wireless networks provides an easy way to make open communication platforms utilizing built-in IQMESH features [25]. This concept is described in details in paper [26]. It will be used as the basis of the concept of IQRF Smart House, building the highest application level and bringing it as completely open platform.

Network functionality of the IQRF platform is based on patented IQMESH protocol. This protocol was defined as a light and portable to the inexpensive hardware with limited resources. IQMESH protocol is scalable and ready to support new routing algorithms. All currently supported routing schemes are ported to the smallest 8b microcontrollers.

To allow integration to other wired and wireless communication systems different gateways exist. IQRF also offer development tools for all products.

ACKNOWLEDGMENT

This research has been supported by the Czech Ministry of Education, Youth and Sports in the frame of MSM 0021630503 *MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies* Research Project, partly supported by ARTEMIS JU in Project No. 100205 *Process Oriented Electronic Control Units for Electric Vehicles Developed on a multi-system real-time embedded platform* and by ENIAC JU in Project No. 120001 *Nanoelectronics for an Energy Efficient Electrical Car*, partly by the Czech Ministry of Industry and Trade in projects FR-TI1/057 *Automatic stocktaking system* and FR-TI1/058 *Intelligent house-open platform*.

REFERENCES

[1] L. De Naris and M. G. Di Benedetto, "Overview of the IEEE 802.15.4/4a standards for low data rate wireless personal data networks.," in 4th Workshop on Positioning, Navigation and Communication 2007 (WPNC 07), 2007, pp. 285-289.

[2] ZigBee. (2009, May) ZigBee Aliance Web Pages. [Online]. http://www.zigbee.org [Cited: 20.9.2010]

[3] C Evans-Pughe, "Bzzzz zzz [ZigBee wireless standard]," IEE Review, pp. 28-31, March 2003.

[4] Khusvinder Gill, Shuang-Hua Yang, Fang Yao, and Xin Lu, "A ZigBee-Based Home Automation System," IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, pp. 422-430, May 2009.

[5] K. Gill, Shuang-Hua Yang, Fang Yao, and Xin Lu, "A zigbee-based home automation system," Consumer Electronics, IEEE Transactions on, pp. 422-430, March 2009.

[6] D Edgan, "The emergence of ZigBee in building automation and industrial control," Computing & Control Engineering Journal, pp. 14-19, April-May 2005.

[7] I.A. Zualkernan, A.R. Al-Ali, M.A. Jabbar, I. Zabalawi, and A. Wasfy, "InfoPods: Zigbee-based remote information monitoring devices for smart-homes," Consumer Electronics, IEEE Transactions on, pp. 1221-1226, August 2009.

[8] R. Casas, A. Marco, I. Plaza, Y. Garrido, and J. Falco, "ZigBee-based alarm system for pervasive healthcare in rural areas," Communications, IET , pp. 208-214, February 2008.

[9] I. Poole, "What exactly is... ZigBee?," Communications Engineer , pp. 44-45, August-September 2004.

[10] T. Ciardiello, "Wireless communications for industrial control and monitoring," Computing & Control Engineering Journal , pp. 12-13, April-May 2005.

[11] Carles Gomez and Josep Paradells, "Wireless Home Automation Networks: A Survey of Architectures and Technologies," IEEE COMMUNICATIONS MAGAZINE, vol. 48, no. 6, pp. 92-101, June 2010.

[12] D. Flowers and Y. Yang, "MiWi Wireless Networking Protocol Stack," 2008.

[13] Y. K. Huang et al., "An Integrated Deployment Tool for ZigBee-based Wireless Sensor Networks," in 5th International Conference on Embedded and Ubiquitous Computing, Shanghai, 2008, pp. 309-315.

[14] T. W. Song and C. S Yang, "A Connectivity Improving Mechanism for ZigBee Wireless Sensor Networks," in 5th International

 ISBN:978-1-61208-113-7

Conference on Embedded and Ubiquitous Computing, Shanghai, pp. 495-500.

[15] C. Gomez and J. Paradells, "Wireless home automation networks: A survey of architectures and technologies," Communications Magazine, IEEE , pp. 92-101, June 2010.

[16] J. Walko, "Home Control," Computing & Control Engineering Journal , pp. 16-19, October-November 2009.

[17] G Ferrari, P Medagliani, S Di Piazza, and M Martalo, "Wireless sensor networks: Performance analysis in indoor scenarios," Eurasip Journal on Wireless Communications and Networking, 2007.

[18] M. H. F. Ghazvini, M. Vahabi, M. F. A. Rasid, and R. Abdullah, "Improvement of MAC Performance for Wireless Sensor Networks," in 13th International-Computer-Society-of-Iran-Computer Conference, 208, pp. 147-152.

[19] L. L. Liang, L. F. Huang, X. Y. Jiang, and Y. Yao, "Design and Implementation of Wireless Smart-home Sensor Network Based on ZigBee Protocol.," in International Conference on communications, Circuits and Systems, Xiamen City, 2008, pp. 487-491.

[20] V.C. Gungor and G.P. Hancke, "Industrial Wireless Sensor Networks: Challenges, Design Principles, and Technical Approaches," Industrial Electronics, IEEE Transactions on, pp. 4258-4265, October 2009.

[21] Microrisc. (2009, May) Microrisc Web Page. [Online]. http://www.microrisc.cz/new/weben/index.php [Cited: 18.5.2010]

[22] V. Šulc, "Czech Republic Patent - Electronic transceiver module for network wireless communication in electric or electronic devices or systems.," PUV 16181.

[23] V. Šulc, "Czech Republic Patent - Module for wireless communication between electric or electronic equipment or systems, method for its control and method for creating generic platforms for user applications in area of wireless communications with those mo," PUV 18340.

[24] V. Šulc, "US Patent - Method of coding and/or decoding binary data for wireless transmission, particularly for radio transmitted data, and equipment for implementing this method.," 7167111, 2007.

[25] V. Šulc, "Czech Republic Patent - A method of accessing the peripherals of a communication device in a wireless network of those communication devices, a communication device to implement that method and a method of creating generic network communication," PUV 18679, 2008.

[26] V. Šulc, R. Kuchta, and Radimír Vrba, "IQMESH implementation in IQRF wireless communication platform," In 2009 Second International Conference on Advances in Mesh Networks, pp. Pages 62-65, 2009.

 ISBN:978-1-61208-113-7

Secure Packet Transfer in Wireless Sensor Networks – A Trust-based Approach

Yenumula B. Reddy
Grambling State University
Grambling, LA 71245, USA
ybreddy@gram.edu

Rastko Selmic
Louisiana Tech University
Ruston, LA 71270, USA
rselmic@latech.edu

***Abstract*—Trust is very important in wireless sensor networks to transfer the data from source to destination. The Dynamic Source Protocol calculates the alternate path, if any node fails to transfer the data. The Dynamic Source Protocol does not have any built-in functionality to calculate an alternate path if the path has a malicious node. With the expense of an intruder detection system we can detect the malicious node and alter the data/packet transfer path. However, intruder detection system is very expensive for wireless sensor networks and there is no guarantee in detecting a malicious node. In the current research a trust-based approach is recommended to minimize the overheads of intruder detection system and it also detects the abnormal behavior nodes. The proposed model uses the repeated games to detect faulty nodes through the cooperative effort in the sensor network and further judges the trust of successive nodes. Simulations were presented for normalized payoff of packet dropping, average discount payoff, and trust relation.**

Keywords-wireless sensor networks; repeated games; packet transfer; trust-based approach; secure transferof data.

I. INTRODUCTION

Wireless sensor networks (WSN) are used in a variety of applications including structural health monitoring (SHM), industrial automation (IA), civil structure monitoring (CSM), military surveillance (MS), and monitoring the biologically hazardous places (BHP). In CSM, MS, and BHP the data is transferred over a number of nodes and any malicious node in the path leads to a dangerous situation. The Dynamic Source Protocol (DSR) cannot detect the malicious node and the IDS package has overheads as well as more false alarms. Hence, we need an alternative approach to detect the malicious node on the communication path with minimum overheads. The alternative approach includes trusting the next node in the path generated by DSR. Here, trust means transferring the packets above expected percentage (for example more than 95%) of packets that were received by that node.

The sinkhole detection, selective forwarding attacks, acknowledgement spoofing, detection of malicious node, and utility-based decision making were discussed in [1-4, 15-19, 21-22]. None of these researchers attempted to verify that the next node in the path was malicious or trustworthy to transfer the data. Failure to transfer the packets depends upon the normal failure of node (communication path or battery loss or node was destroyed) or if the node is compromised. The research of selective forward attacks and detection of malicious nodes provides an extra effort if the data does not reach the destination. But we need a trusted path at the time of transferring the data (packets).

Perrig et al. [1] introduced the modified TESLA [2] protocol for sensor networks and named it µTESLA. The new protocol (µTESLA) is designed to show that security is possible in sensor networks by usage of a simple model to authenticate and transfer the data that is required. Therefore, it is necessary to develop a simple model that eliminates unnecessary checks, avoids sinkholes, detect selective forward packet drops, and improve processing time. The checkpoint-based multi-hop acknowledgement scheme (CHEMAS) [3] identifies the localization of the suspected node that requires extra processing to detect a malicious node. The authors claim that the scheme (CHEMAS) has a high detection rate with communication overhead.

Isolating misbehavior and stabling trust routing in wireless sensor networks was studied in [4]. The trust routing algorithm uses the µTESLA scheme to form the chain of trust. The chain of trust is an expensive process and has more overheads compared to trusting the next successive node. However, it is difficult to keep track of the complete communication path particularly in WSN. The authors in [4] discussed various search methods to detect the insecure locations and isolate those locations from communication paths.

Zhang and Huang [5] used reinforcement learning to establish a secure path for packet transfer from source to base-station. They concluded that adaptive spanning trees can maintain the best connectivity for transferring the packets between source and destination. The authors further discussed the energy-aware and congestion-aware problems for successful delivery of packets.

Carmen et al. [13] discussed the trust management in wireless sensor networks. A trust management system helps to detect the node (faulty or malicious) behaving in an unexpected way. Liu et al. [23] presented a dynamic trust model for ad hoc networks, where each node is

 ISBN:978-1-61208-113-7

assigned a trust value according to its identity. Sometimes trust level is also calculated by evaluation of nodes over other nodes. Evaluation of trust factor is done with IDS data and statistical data of packet transfer rate. Rebahi et al. [9] discussed a reputation based trust mechanism in ad hoc networks, where each node monitors the neighboring nodes activities, sends the information to the reputation manager, and stores it in a matrix for evaluation of nodes.

The belief-based packet forwarding model in mobile networks using repeated games was discussed in [6]. The authors described the belief-based packet forwarding model as being dependent upon past history of other nodes' information transfer. The model enforces cooperation in the ad hoc networks with noise and imperfect observation. Enforcing the cooperation slightly degrades the performance of packet transfer compared to unconditionally cooperative outcomes. The model further provides the ad hoc networks and needs to modify for WSN.

The rest of the paper introduces the repeated games to model the trust level of successive node and then formulate the trust-based model in a cooperative environment. Further, we calculate the trust-based packet forwarding and discuss the future research.

II. TRUST MANAGEMENT

Trust is subjective term used for reliability of an entity. It is a subjective probability of an individual A expects another individual B to perform a given task. The trust management model helps to detect the intruders (malicious nodes) and discard them from the communication path [9, 11, 12, 13]. The concept of reputation (collecting the data about status of a successive node) linked to trustworthiness [10] depends upon trusting a person (node). In the current situation trust depends upon the ratings of successive the node. If the ratings of the successive node are above the expected value (threshold) then the node will be trusted for transfer of data. Further, relaying on self detecting misbehavior nodes (intruders) is dangerous and collaborating between neighboring nodes is required.

Figure 1 shows the data transfer scenario from node A through node D and establishing the trust of node D for future data transfer. For example, node A sends data to node D and node D receives the data and acknowledges to node A. There is no guarantee that node D transfers the data to the next node in the path. If node A knows that node D transferred the data successfully, then node A assumes that node D can be trusted. After repeated transfers (successive node activity), if the trust factor reaches below the threshold, then node A compares the trust factors of its neighboring node B and node C that are transferring their data through node D. If nodes B and C trust node D, then node A establishes a new route for successful transfer of data and avoids node D. Trust of the next successive node in data path is a kind of watchdog approach to detect the malicious node.

In the proposed approach, each node maintains a rating of its successive node (number of successful pack transfer) in the path. If the ratings of a successive node are above the threshold (minimum error rate) then the current node continues to transfer the packets. The current approach does not expect to calculate all ratings (packet transfer, noise, jamming, and infection factor) of its neighboring nodes and selects the path of highest ratings [1]. Selecting a highest rating path requires more processing time and is a waste of energy in the sensor node. The proposed approach detects the malicious node using the trust factor. For example, if node D only selectively drops the packets from node A but not from nodes C and D then node A concludes that the path from node A through node D cannot be trusted and node A establishes the alternative path. The alternate path is selected only if the successive node is not trusted.

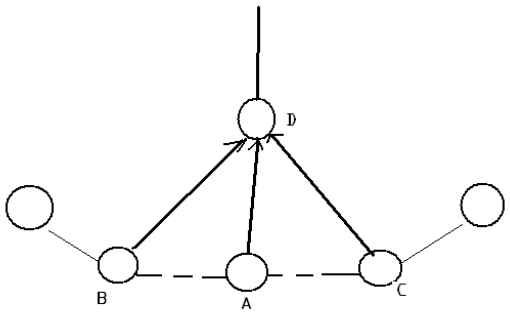


Figure 1. Scenario for node 'A' establishing trust of node 'D'.

III. GAME MODEL

In games [8, 20] the interaction between the players is inherently dynamic, so players always observe the actions of other players and decide their optimal response. Many times, the game is played repeatedly and decisions depend upon the previous actions or conclusion of previous actions. In repeated games, players have more opportunity to learn to coordinate their actions depending upon the previous outcome. In Figure 1, Player 1 and Player 2 (node A and node D) are involved in transferring the information where Player 1 transfers data to Player 2. Player 1 then waits for successful transfer of data packets from Player 2 to the next step in the path. Player 1's trust on Player 2 depends upon Player 2's successful transfer of data packets. The problem is how these two players coordinate their actions.

The outcome of Player 1 depends upon the actions (repeated outcome conclusion) of Player 2. In the cooperative effort, we must consider the outcome of neighboring players (within communication distance) of Player 1; that is, Player 3 and Player 4 (node B and node C in Figure 1) and have the similar interaction with Player 2. If the outcomes of Player 3 and Player 4 are the same as Player 1 (no better than Player 1) then the Player 1

 ISBN:978-1-61208-113-7

concludes either to transfer the future packets or chooses an alternative path. If the trust relation of Player 1on Player 2 is consistent and depends upon the outcome of its neighbors then we say it reaches to Pareto optimality.

In repeated games the behavior of Player 1 depends upon its opponent's (Player 2) actions (behavior). Further, no threat, punishment, or revenge is considered. The strategy is that Player 2 must transfer the packets received from Player 1. The trigger strategy is that the malicious behavior of Player 2 will permanently disconnect the path from Player 1 and its neighbors that have the current path through player 2. For example, the Stage game G is of the form

$$G = (N, A, U) \quad (1)$$

where N is a set of users (set of sensor nodes), A is a set of pure strategy profiles (actions – action may be the missing packets for each transmission), and U is a vector of payoffs. If Ω is the common discount payoff and $g_i(a^t)$ is the per-period payoff of the i[th] node related to current action a^t, then the normalized payoff β (relation to utility of sequence $a^0, a^1, .., a^T$) at any node is given by [20]

$$\beta = \frac{1-\Omega}{1-\Omega^{T+1}} t \sum_{t=0}^{t=T} g_i(a^t) \quad (2)$$

The trust of the player depends upon the outcome of β. The Figure 2 shows that the payoff is higher with a lower number of packets dropped in the same time period. But the average payoff will be very close in a large time period. Therefore it is necessary to consider frequent averages for packet dropping for appropriate decision.

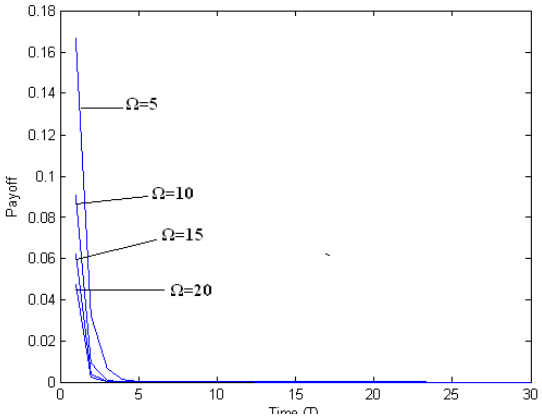


Figure 2. Payoff β verses packet dropping in a given time period

IV. Trust Model and Game Application

Each node in the sensor network maintains a dynamic table to store the information about packet transfer of the successive node in the path. The values in the table include the packets transmitted from the node and packets transferred from the successive node (recorded through over hearing). These values are used for trust calculation of the successive node. The values are also used to calculate the risk involved in order to carry out packet transfer. In other words trust value is a simple mathematical representation. The problem with no successive node will be dealt with different models [14, 15].

Consider a sensor network of N nodes deployed in a field. Let the nodes be connected as shown in the Figure 3 and represented through a matrix of equation (3). The filled nodes are existing nodes and unfilled are drawn to complete the matrix. Unfilled means no node exists or a dead node. The equation (3) helps to verify the isolated node (blackhole).

$$M = [M_{i,j}] = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \quad (3)$$

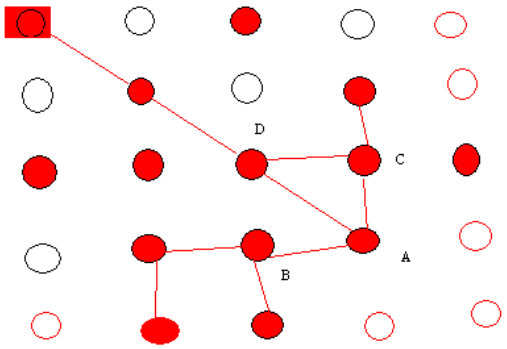


Figure 3. Sensor Network Nodes and their relation with neighboring nodes

Reputation is used to predict the behavior of the node. We create a table at node i (values stored in table at node i are over hearing from Node 2) to predict the behavior of the node j. Let $R_{i,j}$ represents the reputation of node j represented by node i. The reputation table RT_i stores the reputations maintained by node i and represented as:

$$RT_i = \{R_{i,j}\} \quad (4)$$

The periodic quantification of reputations at node j is $Q_{i,j}$ and is stored at RT_i as part of node j. The missing is calculated as (1 - $Q_{i,j}$). Further, each node has direct and indirect observations of reputations. Direct observation is the reputations stored at node i and indirect observations are received from neighboring node (s). The indirect

observations are represented as $IQ_{i,j}$. The trust prediction of the node j depends upon the $Q_{i,j}$ and $IQ_{i,j}$.

In repeated games, expected payoff depends upon the action profile and its observation. The action profile is given by

$$U_i = (\frac{1}{Q_{i,j}})\lambda \quad (5)$$

where λ is the difference between $Q_{i,j}$ and $IQ_{i,j}$. If λ =0 then the packets transferred at a node and its neighbor node is the same. The trust of the node depends upon the factor β. Further we calculate the average discount factor to calculate the stable state of the node. The average discount payoff is given by

$$UA_i = \beta(\sum_{t=1,n} \Omega_i(t).U_i(t)) / n \quad (6)$$

If the average discount payoff is above the threshold then node is in trust state and if trust state is consistent then we say it reaches Nash equilibrium. If the Nash Equilibrium exists in repeated games, then it satisfies Folk theorem [7] and sufficiently the player reaches to Pareto optimal payoff in Nash equilibrium. The simulations for average discount payoff are shown in Figure 4.

For a small value of λ (0.001) and probability of more than 90% successful packet transfer rate, the payoff increases in a smaller period of time (if lower number of packets is dropped). In average discount payoff, the number of packets dropped is set approximately the same. The number of packets transmitted is numbered in small or many. The average discount pay of increases initially (from 100 packet transmission to 900 packet transmission) and settles after it reaches a transmission rate of 1000 packets with the same number of drops. This shows, for a selected action strategy of a player, the game reaches Nash equilibrium at action profile during the time period of higher number of packet transmission with lower dropouts. That means the successive node can be trusted at current state.

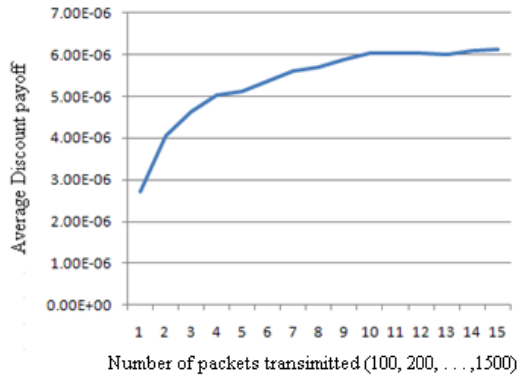


Figure 4. Average discount payoff verses number of packets dropped

V. TRUST-BASED PACKET FORWARDING

In trust-based systems, we begin to believe all nodes in the path are trusted. Trust of node 2 at node 1 will be developed after repeated transfer of packets from node 1 (n_i) to node 2 (n_j) and then successfully transferred from node 2. The trust of interaction between these nodes is

$$T_{i,j}^t = (n_j, s_k, TE_{i,j,t}) \quad (7)$$

where $T_{i,j}^t$ is a trust of node n_i on node n_j at time t , s_k is a set of possible specifications to perform task at n_j where $s_k \in S$, and $TE_{i,j,t}$ is the set of tasks.

Further, the node n_i , the initiator node must store the data about the reliability of node n_j when the packets are transferred repeatedly. The node n_i experience in repeated operation of packet transfer is

$$R_{i,j}^t = (n_j, s_k, P_{i,j,t}) \quad (8)$$

where $P_{i,j,t}$ is satisfaction achieved by node n_i at node n_j at any time t and $P_{i,j,t} \in (0,1)$.

The experience of each particular task will be updated at n_i and represented as

$$I^t(n_j, s_k) = (n_j, w_j) \quad (9)$$

where w_j is the response from n_j in the interaction. By updating the process combinations of I^t and storing the experiences of $T_{i,j}^t$ and $R_{i,j}^t$ we get the quality satisfaction measurements.

The equations (2), (6), and (9) will provide the needed information to trust the node n_i for future transformation of information.

To create trust level we generated random data to test the equation (9). In the test process, 100 random samples were generated for node n_j . If node n_j is trusted more than 90%, we note that the trust level is above threshold. This process was repeated 100 times to reach correct trust level. The process was repeated and the percentage of trust in hundred attempts is shown in Figure 5.

The random generation of trust data is not a correct process but it helps in simulations. The average trust of a hundred samples in Figure 5 is approximately 90.42. The average hundred samples each time is approximately 90.42. The threshold was set as 90 and above and satisfies the simulation results. Therefore, we can assume that if the transfer rate is above 90% the node can be trusted.

 ISBN:978-1-61208-113-7

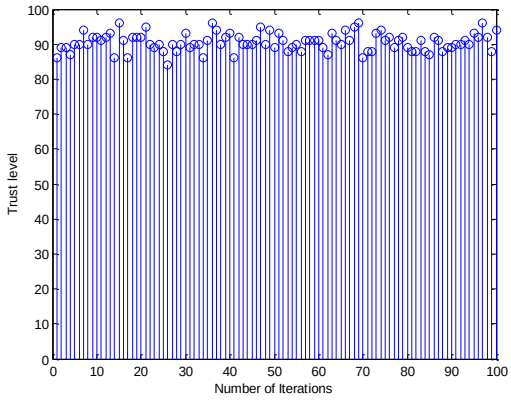


Figure 5. Trust relation generated in 100 iterations

VI. TRUST REPUTATION AND INTERACTING WITH NEIGHBORS

To confirm the trust of the successive node the node interacts with its neighbors. The neighbors of node n_i can be represented as:

$$N_i = \{n_k \mid n_k \in N, then eighbor(n_i, n_k) = true$$

In Figure 1, nodes B and C are neighbors of node A, if the Boolean function value is true. Similarly, the current node A interacts with several of its neighbors to create trusted neighbors and keeps the superior nodes and ignores the inferior nodes. For example, if we denote ζ_i as the inferior neighbor node and ζ_s as the superior neighbor node then their values will vary as $0 \leq \zeta_i \leq \zeta_s \leq 1$. For the stronger neighbor, the relational value must be close to 1. Therefore, the representation of most trusted node is

$$NT^t_{\sup}(n_i, s_k) = \{n_k \mid n_k \in N, \quad if\ trust\ of\ \ n_k \geq threshold \qquad (10)$$

Similarly, the set of nodes with doubtful confidence is given by

$$NT^t_{\inf}(n_i, s_k) = \{n_k \mid n_k \in N, \quad if\ trust\ of\ \ n_k < threshold \qquad (11)$$

The most reputed nodes (established complete trust over time) will be grouped into reliable nodes and represented as

$$NR^t_{\sup}(n_i, s_k) = \{n_k \mid n_k \in N, \quad if\ trust\ of\ \ n_k \geq threshold \qquad (12)$$

The reliable nodes will be used as a reference to verify the trust of successive nodes. If the reliable node is not available, it will verify with a trusted node before it transfers the packets.

The calculation of the threshold value is very important and will be calculated using equation (8). The threshold value will be updated in preset timings by the agent.

VII. CONCLUSIONS AND FUTURE RESEARCH

The current available research models deal with secure transfer of packets, intruder detection, sinkholes, and similar approaches. All these methods need a lot of processing, storage, and energy. There is no literature available for a simple security model for wireless sensor networks that confirms the successive node to transfer the packets. The proposed model is a unique approach to transfer the data securely and at the same time confirms the trust of next level nodes. We are working on the following research ideas that transfer the packets securely from source to destination.

a) What happens if an intruder at successive node level acts as a real node and acknowledges to the preceding node with 100% success of packet transfer and then transfers the packets to the sinkhole?
   - *This problem was solved using the NS2 package by creating a table at the previous node and observing the successive node. The experiment will be useful for detecting the sinkhole. The results will be presented in the next conference.*

b) What happens if the intruder modifies the packets and forwards them to the next level and then these corrupted packets reach the destination?
   - *This is an open problem and will be attempted and solved soon.*

c) What happens if the intruder stores the packet forwarding table appropriately (as the preceding node requires for successful transformation) and never forwards the packets (acts as an intelligent sinkhole).
   - *This problem will be solved with (a) before we publish the results.*

We are working on the above problems by modifying the node level code of the NS 2 package. In the first step, a large size sensor network with 1000 nodes was created and experienced heavy dropping of packets due to overloading at the node. We then minimize the size of the network to 500, 400, 300, 200, and 100 sensor nodes and succeeded partial control of dropping the packets. So, we decided to start with less than 25 nodes for simulations and the packet dropping was controlled. Further, the proposed model is more realistic compared to the previous models in the research [4, 5, 6] and is simple to implement.

ACKNOWLEDGEMENT

The research work was supported by the ONR with award No. N00014-08-1-0856. The first author wishes to express

 ISBN:978-1-61208-113-7

appreciation to Dr. Connie Walton, Grambling State University and Dr. S. S. Iyengar, LSU Baton Rouge for their continuous support.

REFERENCES

[1] Perrig, A., Szewczyk, R., Wen, V., Culler, D., and Tygar, J. D., "SPINS: Security Protocols for Sensor Networks", MOBICOM 2001, Rome, Italy, June 2001.

[2] Perrig, A., Canetti, R., Tygar, J. D., and Song, D., "Efficient authentication and signing of multicast streams over lossy channels", IEEE Symposium on Security and Privacy, May 2000.

[3] Xiao, B., Yu, B., and Gao, C., "CHEMAS: Identify Suspect Nodes in Selective Forwarding Attacks", Journal of Parallel Distributed Computing, Vol 67, 2007.

[4] Tanachaiwiwat, S., Dave, P., Bhindwale, R., and Helmy, A., "Location-centric Isolation of Misbehavior and Trust Routing in Energy-constrained Sensor Networks", IEEE IPCC, October 2004.

[5] Zhang, Y., and Huang, Q., "A Learning-based Adaptive Routing Tree for Wireless Sensor Networks", J. of Communications, 1 (2), 2006.

[6] Ji, Z., Yu, W., and Liu, K. J., "Belief-based Packet Forwarding in Self-organized Mobile Ad Hoc Networks with Noise and Imperfect Observation", IEEE WCNC 2006.

[7] Abreu, D., Dutta, P., and Smith, L., "The Folk Theorem for Repeated Games: A NEU Condition", Econometrica, Vol. 62, 1996.

[8] Yuan, J., and Yu, W., "Distributed cross-layer optimization of wireless sensor networks: a game theoretic approach", Proc. of IEEE Global Telecommunications Conference, 2006.

[9] Yacine R., Vicente E., Mujica V., and Dorgham Sisalem., "A Reputation-Based Trust Mechanism for Ad Hoc Networks", 10th IEEE Symposium on Computers and Communications (ISCC'05), 2005.

[10] Audun J., Roslan I., and Colin B., "A survey of Trust and Reputation Systems for Online Service Provision", Decision Support Systems, 2006.

[11] Mohammad M., and Subhash C., "Trust management in Wireless Sensor Networks", 5th IEEE/ACM international conference on Hardware/software codes and system synthesis, 2007.

[12] Junbeom H., Yoonho L., Seongmin H., and Hyunsoo, Y., "Trust-based secure aggregation in Wireless Sensor Networks", Sensor and Ad Hoc Communications and Networks (SECON '06), 2006.

[13] Fernandez-Gago, M. C., Rodrigo, and R., Javier L., "A Survey on the Applicability of Trust Management Systems for Wireless Sensor Networks", 3rd International workshop on Security, Privacy, and Trust in Parvasive and Ubiquitous Computing, July 2007.

[14] Reddy, Y. B., "Potential Game Model to Detect Holes in Sensor Networks", IFIP/NTMS 2009.

[15] Kanno, J., Buchart, J. G., Selmic, R. R., and Phoha, V., "Detecting coverage holes in wireless sensor networks," 17th Mediterranean Conference on Control and Automation, June, 2009.

[16] Mark F., Jean-Pierre H., and Levente B., "Cooperative Packet Forwarding in Multi-Domain Sensor Networks", PERCOM 2005.

[17] Garth V. C., and Niki P., "Evolution of Cooperation in Multi-Class Wireless Sensor Networks", LCN 2007.

[18] Narayanan, S., Mitali S., and Bhaskar K., "Decentralized utility-based sensor network design", Mobile Networks and Applications, June 2006.

[19] Kannan, R., and Iyengar, S.S., "Game-theoretic models for reliable path-length and energy-constrained routing with data aggregation in wireless sensor networks", IEEE J. of selected areas in communications, Aug 2004

[20] Machado, R., and Tekinay, S., " A survey of game-theoretic approaches in wireless sensor networks", Comput. Netw. 52, 16, Nov. 2008.

[21] John, B., and Gabriel, N., "Utility-based decision-making in wireless sensor networks", Proc. of the 1st ACM international symposium on Mobile ad hoc networking & computing, November 20, 2000.

[22] Miller, D.A., Tilak, S., and Fountain, T., "Token equilibria in sensor networks with multiple sponsors", Collaborative Computing: Networking, Applications and Worksharing, 2005.

[23] Zhaoyu L., Anthony W. Joy., and Robert A. T., "A Dynamic Trust Model for Mobile Ad Hoc Networks", IEEE International workshop on Future Trends of Distributed Computing Systems (FTDCS) 2004.

 ISBN:978-1-61208-113-7

Opportunistic Sensing in Wireless Sensor Networks

Hans Scholten and Pascal Bakker
Pervasive Systems
University of Twente
Enschede, the Netherlands
hans.scholten@utwente.nl, aboe@aboe.nl

***Abstract*—Opportunistic sensing systems consist of changing constellations of wireless sensor nodes that, for a limited amount of time, work together to achieve a common goal. Such constellations are self-organizing and come into being spontaneously. This paper presents an opportunistic sensing system to select a subset of sensor nodes out of a larger set based on a common context. We show that it is possible to use a wireless sensor network to make a distinction between carriages from different trains. The common context in this case is acceleration, which is used to select a subset of carriages that belong to the same train. Simulations based on a realistic set of sensor data establish that the method is valid, but that the algorithm is too complex for implementation. Downscaling reduces the number of processor execution cycles as well as memory usage, and makes the algorithm suitable for implementation on a wireless sensor node with acceptable loss of precision. Actual implementation on wireless sensor nodes confirms the results obtained with the simulations.**

***Keywords*-opportunistic sensing, wireless sensor network, context awareness, activity recognition**

I. Introduction

"Opportunistic sensing is seen as a way to gather information about the physical world in the absence of a stable and permanent networking infrastructure." (Opportunity Workshop at Ubicomp 2010, Copenhagen, Denmark). The absence of a stable and permanent networking infrastructure dictates that collected information is either processed and acted upon inside the network by opportunistic collections or clusters of nodes [1], or the information is preproceesed and stored inside the network untill there is an opportunity to forward it outside the network, as is the case in delay tolerant networks [2], [3], [4], [5].

Opportunistic sensing or networking is often associated with human-centric ubiquitous systems, such as in crowd sourcing and participatory sensing applications [6], [7], [8] or are focusing on human activity recognition [9], [10], [11].

In this paper we show how context awareness based on a common pattern of movement is used to select only those carriages from a much larger population that belong to the same train. Movement as a discriminating factor for context awareness has been described earlier [12], [13], but not for trains. The problem was first introduced in [14] describing a communication protocol for wireless networks in linear structures such as trains. The network is part of a railway safety system to monitor the initial composition of a train and to detect any change in composition once the initial composition is established. In Europe many different safety systems exist and trains crossing borders must be equipped with all applicable safety systems. These systems are infrastructure based and integrated in the tracks, whereas the proposed system is entirely train based. Composing a train not only takes place on switchyards, but also on the way when carriages are added to the train or decoupled from it. Although a simple beacon based detection system seems appropriate, it would not suffice to check the train's composition. Because the ID of the beacons in sight are not known a priori, nor the mapping of IDs to carriages, nor carriages to trains, there is no way to discriminate carriages in different trains in close proximity. Additional measures must be present to select a train's beacons from all beacons that are in communication range. In this paper, motion is used as the discriminating feature between trains.

In the remainder of this paper we will discuss the collection and analysis of the data sets to be used in the simulations, the simulation itself and the implementation on wireless sensor nodes respectively, followed by a discussion of the results.

II. Data Collection and Analysis

As explained in the introduction, movement is used as discriminating feature. However, motion is multi-dimensional and too complex to use in wireless sensor nodes. Not only processing power and memory usage might be issues, also running complex algorithms for longer periods of time consumes large amount of energy, negatively influencing the operational time of the system. Two measures are taken to enable implementation on nodes:

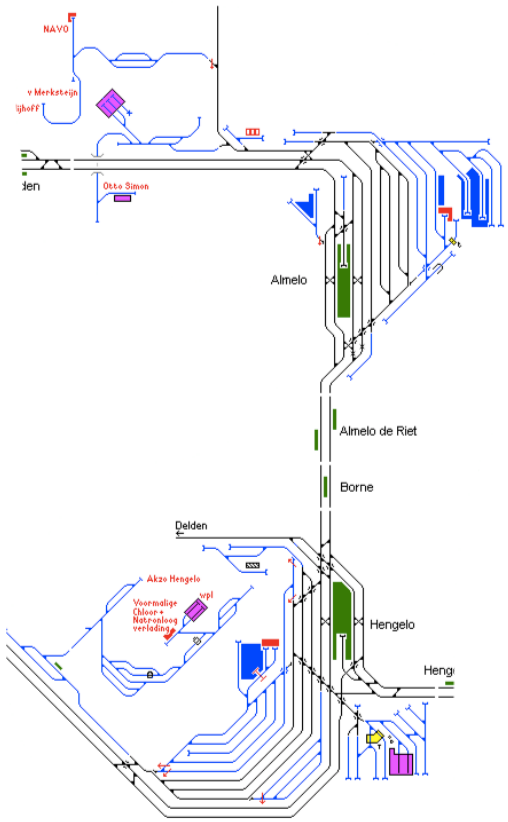
- limit the time the algorithm executes, and
- simplify the input data for the algorithm.

Under normal conditions a train's composition will not change while underway and moving. If a change occurs under these circumstances, it will be accidentaly. The algorithm can be split into two phases:

- when the train starts moving, detect the initial set of carriages that belong to the train, and
- once the train is moving, only check those carriages in the initial set and ignore all others.

 ISBN:978-1-61208-113-7

(a) Wireless sensor node



(b) Railway track

Figure 1. Data collection

The latter phase is accomplished by periodically pinging the initial set, which is a simple process that consumes less energy than is needed for the first phase.

A. Collecting the Data

To simplify phase 1, a train's movement is analysed to reveal those characteristics that are essential and those that can be safely ignored. Representative data sets are recorded on a track featuring multiple stops and curves (see Figure 1b), resulting in a wide variety of data. Figure 1a shows the wireless sensor node that is used to sample the data, consisting of an Ambient muNode 2.0 provisioned with an STMicroelectronics LIS3LV02DQ accelerometer. The maximum sample rate of this sensor is 640 Hz, but the used combination of hardware and software gives a maximum sample rate of 160 Hz. The sensor nodes are aligned with the horizontal x-axis in the driving direction, the y-axis in the horizontal sideways direction and the z-axis in the vertical direction.

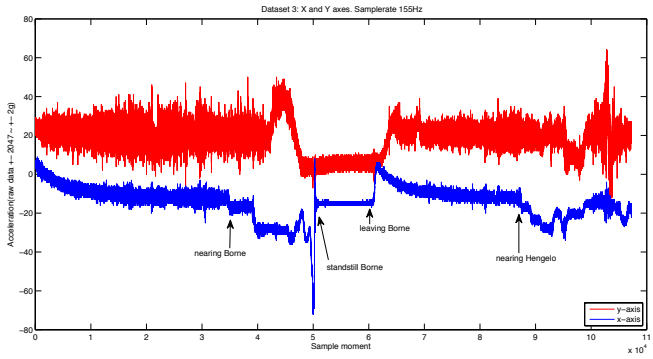
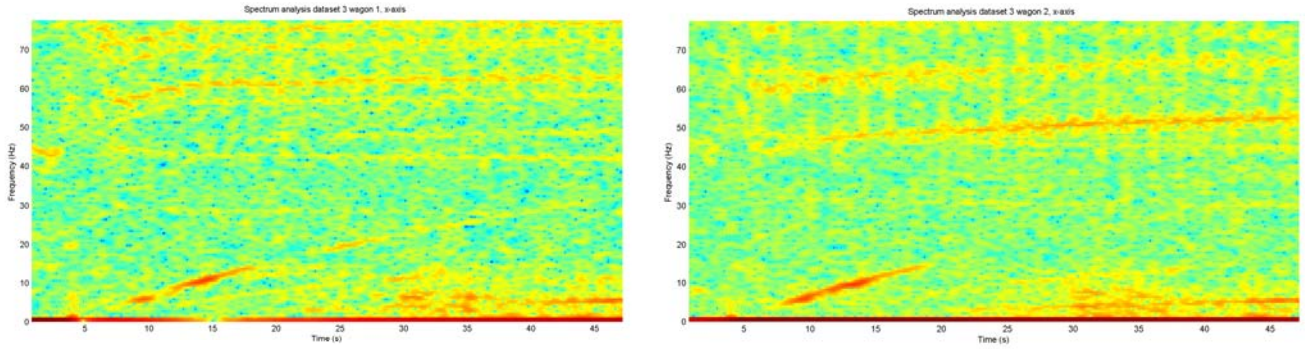


Figure 2. Accelerometer x- and y-axis

B. Analysing the Data

Figure 2 shows raw data with a sampling frequency of 155 Hz of a journey between two stations with one



(a) Spectrum carriage 1, train 1, x-axis

(b) Spectrum carriage 2, train 1, x-axis

Figure 3. Frequency spectrum of two carriages in the same train

station in between. The bottom graph shows the x-axis and halfway the train is standing still (no acceleration). The top graph depicts the sensor's y-axis and just before reaching the middle station the train is crossing a switch changing tracks. The data from both x- and y-axis (and, though not shown here, from the z-axis as well) contain a lot of noise and must be filtered with a high-off filter before it can be useful. Figure 3 shows the frequency spectrum for the x-axis data of two carriages in the same train starting to move. Only frequencies lower than the sampling frequency/2 are considered. The spectra are similar in the low frequencies, but quite different in the high frequencies. The spectra of two carriages in two different trains on the same part of the railway track (not shown) differ in all frequencies, but show similar characteristics in the lower frequencies. Closer inspection learns that a cut-off frequency of 2 Hz can be applied without losing discriminating features.

Data from the y- and z-axis of two carriages are similar irrespective whether they are in the same train or not and do not contain enough discriminating features. In the remainder of the paper only data from the x-axis of the accelerometer are considered.

To filter the data, a second order Butterworth low pass filter is used. This choice is made because this type of filter will run on the wireless sensor nodes. Figure 4 shows the result: the top and bottom graph are from carriages in the same train, while the middle one is in a different train. The graphs are synchronized, i.e., they are shifted in time so they show trains starting at the same time. In practise it will rarely happen that two trains in communication range will accelerate at exactly the same time. The data sampling rate in the original data set is 155 samples per second. Because the data is filtered at 2 Hz, this high sampling rate is overkill and could be reduced significantly in the final implementation. A frequency of 35 samples per second gives the same results as before. We did not test lower sampling rates.

C. Data Correlation

The last step in the algorithm is to check whether two carriages share the same context by way of correlating the data. The correlation process uses a sliding window over which the data is compared. A wider window normally leads to a more precise result, but also takes longer to produce this result. A smaller window gives a better reaction time, but the result is unreliable. So a trade-off has to be made. Figure 5 gives correlation results at a window size varying from 1 to 5 seconds for a time frame of 7000 samples or 45 seconds. Detection of the carriages is only active in phase 1 of the algorithm when the train starts moving, which is approximately during the first 2000 samples (phase 2 only pings known carriages). Therefore the results after 2000 samples will be ignored. For two carriages in the same train a window size of 155 samples or 1 second would already suffice. However for two carriages in different trains the window size should be at least 465 (3 seconds), or even better 620 samples or 4 seconds.

III. Implementation

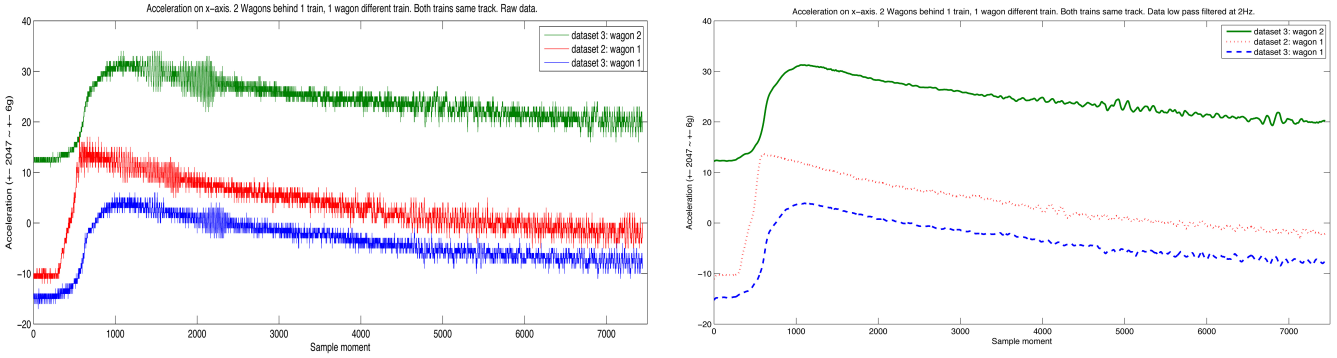
In the following we discuss some implementation details.

A. Optimizations

Matlab calculations, based on realistic data, in the previous section showed that to select a set of carriages that belong to the same train out of a larger population of carriages, movement, or more precise acceleration in the driving direction, can be used as a discriminating factor. However, the Matlab routines must be optimized or simplified before they will run on the wireless sensor nodes.

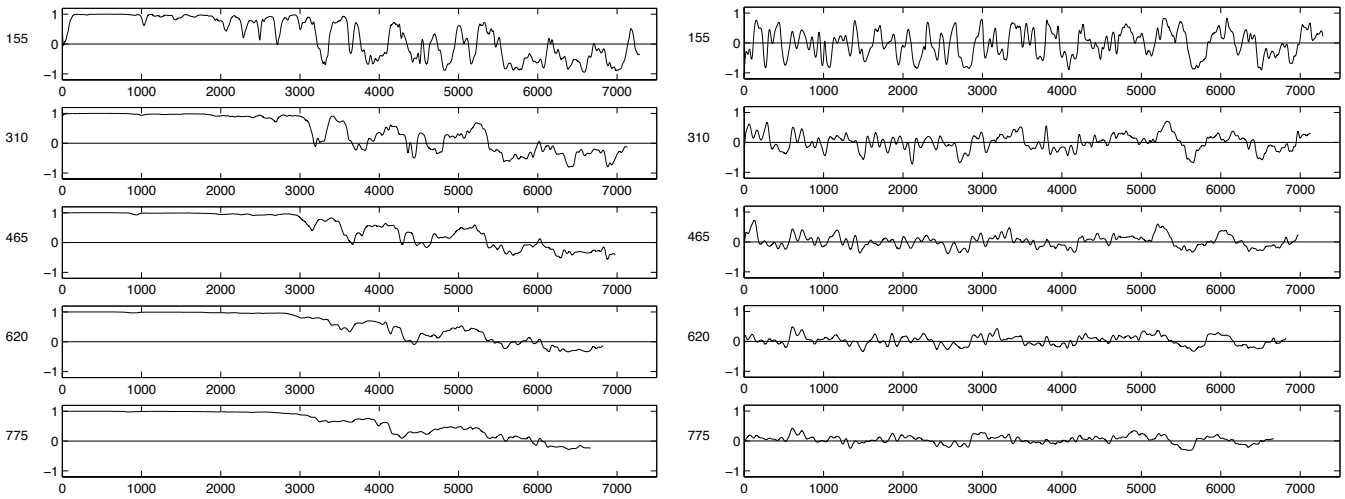
The Matlab program is centralized and assumes that all data is available when needed, which is not the case in the real world. The program and the data are distributed over the wireless nodes: each carriage must calculate its correlation with its neighbors and to do so it needs the movement information from its neighbors. This process is optimized by dynamically forming master/slave pairs. The

 ISBN:978-1-61208-113-7



(a) Unfiltered data (b) Filtered data

Figure 4. X-axis acceleration data from three carriages



(a) two carriages in the same train (b) two carriages in different trains

Figure 5. Correlation of two carriages with varying window sizes

slave sends its data to the master and after calculation the master sends the correlation results back to the slave. For every master/slave pair the movement data is communicated once and the correlation calculations is performed once, thus reducing both needed bandwidth and execution load. To balance the energy consumption on pairs the master and slave switch functionality after each correlation.

One more change in the original Matlab routines must be made before they can be implemented. The filter and correlation from the previous section are based on calculations that use floating point numbers. This puts too much of a burden on the sensor node and a fixed point calculation would be better, though at the cost of possible loss of precision. Errors in rounding results would accumulate and might lead to significant deviations over time. In Figure 6 the difference between floating point and fixed point correlation calculation is shown. The deviation starts to show after 15 to 20 seconds. Since the determination of the train composition takes place in the first 5 to 10 seconds, replacing floating point by fixed point calculations does not influence the end result.

B. Timing and Memory Usage

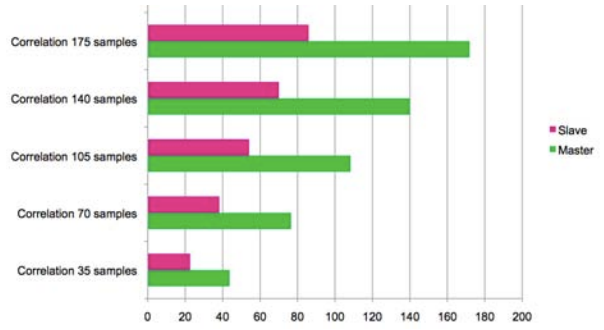


Figure 7. Execution times on a muNode with MSP430 microcontroller

In each master/slave pair, the correlation calculation is

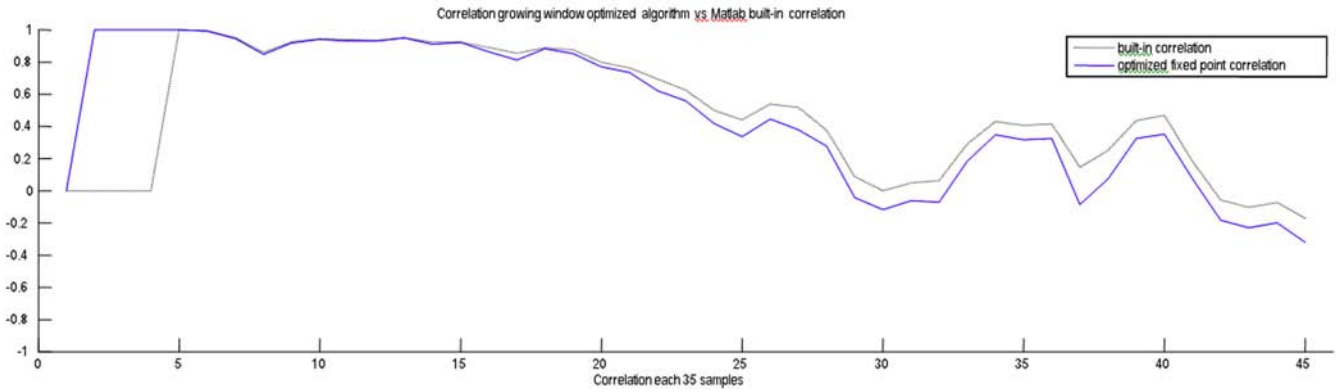


Figure 6. Comparison of floating point and fixed point calculations

done by the master while both partners filter their own data. In addition, the slave needs some time to assemble the packets to be send to the master. In Figure 7 the execution times for filtering and correlation for master and slave are shown. Assuming a correlation window size of 5 seconds (corresponding with 175 samples at 35 Hz), the time needed for the calculations is 171.9 ms and 85.9 ms for master and slave respectively. This leaves room for approximately 6 correlation calculations per second.

Data structure	Size (bytes)
Dataset slave 175 samples	350
Timestamp start	2
Master mean data	4
Slave mean data	4
Master squares data	20
Slave squares data	20
Master sum data	20
Slave sum data	20
Sum products data	20
Total	460

Figure 8. Memory consumption per master/slave pair

The memory consumption of our algorithm on each node depends on the number of master/slave pairs each node has formed with neighboring nodes. The minimum memory consumption for the correlation calculation algorithm on a node is (175+35)*2=420 bytes. Each node stores its own data samples for the given window as well as 35 extra samples, since the point at which master/slave pairs are formed is not the same for each pair. The table in Figure 8 summarizes the memory consumption per master/slave pair. The total amount of bytes is 460 per master/slave pair.

A standard muNode 2.0 has 10kB RAM available. For the normal operation of the node 2kB has been reserved, which leaves 8kB for the correlation algorithm. Given the memory consumption of 460 bytes for a participating node and the availability of 8192 bytes, a node can store up to 16 master/slave pairs in memory.

IV. Conclusion

In this paper we have presented an opportunistic sensor system that makes a selection out of a much larger set based on a common context. In this case, motion information is used to distinguish carriages belonging to different trains. The motion information consists of data obtained by 3-dimensional accelerometers attached to individual train carriages. This data stream then was analyzed with Matlab on a PC to extract the best possible features to discriminate carriages. While sideways motion gives information on track changes and vertical movement indicates the quality of a track, the best information for our purpose is movement in the direction of travelling (x-axis). A spectrum analysis learns that not all frequency components in the sampling data are equally useful: only those below a frequency of around 2 Hz are significant to separate carriages. After filtering, the data from two different carriages are correlated with a correlation window of 5 seconds. We found that a smaller window of approximately 1 second suffices to find two carriages in the same train, but also leads to false positives for carriages in different trains. Extending the window to 5 seconds, no false positives were detected, but this needs further analysis.

After this theoretical confirmation that motion information can be used for context awareness, the algorithm is implemented on wireless sensor nodes. However, the nodes used have several limitations. One limitation is the absence of floating point calculations. First of all the filter and correlation routines are rewritten, so only fixed point calculations are used. This might lead to errors due to accumulation of rounding successive results, but we showed that in the time frame the algorithms run this is not a problem.

The second limitation is power consumption. Filtering and correlation are so computing intensive that they cannot run over longer periods of time without exhausting the battery quickly. A first step is the reduction of the sampling rate

 ISBN:978-1-61208-113-7

from the original 160 Hz to 35 Hz. This is made possible because only the lower frequencies in the sampling data are significant. A lower sampling frequency would have been possible, but this does not substancially contribute to decreasing the processing load. Sampling data, filtering and performing one correlation per second takes 171.9 ms (worst case), which results in a duty cycle of around 17 percent. This exhausts the battery in a couple of hours, at most days, where 6 months is needed. The solution is found by executing the algorithm only for a period of 10 seconds when the train starts moving after each stop. This is enough to establish which carriages belong to one and the same train. During the ride, the initially detected carriages (but not the sequence of the carriages) needs to be confirmed, which can be accomplished by pinging all known carriages at regular intervals.

The last limitation is memory capacity. In our algorithm carriages are correlated in pairs. A carriage is part of as many pairs as it has neighbors. Each pair consumes up to 460 bytes of memory in both partners. With the given memory capacity, a node can accommodate up to 17 neighbors. With a maximum of 6 correlations per second it takes a node 3 seconds to check all its neighbours.

The circumstances in which the data is collected and the implementation is tested form a worst case scenario. The trains that are used in the tests are of the same type with similar characteristics. They all exhibit the same pattern in acceleration and braking, making the data to correlate very similar. We suspect this is the main reason it takes up to 5 seconds to check carriages from different trains (and only 1 second when they are in the same train). This needs further investigation.

Another future research topic is the use of better accelerometers. Those used now measure up to 2g and can be used on trains that accelerate moderately. However, heavy trains that accelerate and brake more slowly have less distinctive movement patterns and need more sensitive sensors.

ACKNOWLEDGMENT

This work is supported by the iLAND Project, ARTEMIS Joint Undertaking Call for proposals ARTEMIS-2008-1, Project contract no. 100026

REFERENCES

[1] H. Scholten, R. Westenberg, and M. Schoemaker, *Sensing train integrity*, IEEE Sensors 2009 Conference, pages 669674, Los Alamitos, October 2009. IEEE Computer Society Press.

[2] Schwartz, R.S. and van Eenennaam, E.M. and Karagiannis, G. and Heijenk, G.J. and Klein Wolterink, W. and Scholten, J, *Using V2V communication to create Over-the-horizon Awareness in multiple-lane highway scenarios*, IEEE Intelligent Vehicles Symposium (IV) 2010, 21-24 June 2010, La Jolla, CA, USA. pp. 998-1005. IEEE Computer Society Press. ISSN 1931-0587 ISBN 978-1-4244-7866-8

[3] M. Kumar, *Distributed computing in opportunistic environments*, UIC 09: Proceedings of the 6th International Conference on Ubiquitous Intelligence and Computing, pages 11, Berlin, Heidelberg, 2009. Springer-Verlag.

[4] L. Lilien, A. Gupta, and Z. Yang, *Opportunistic networks for emergency applications and their standard implementation framework*, Performance, Computing, and Communications Conference, 2002. 21st IEEE International, 0:588593, 2007.

[5] L. Pelusi, A. Passarella, and M. Conti, *Opportunistic networking: data forwarding in disconnected mobile ad hoc networks*, Communications Magazine, IEEE, 44(11):134 141, november 2006.

[6] R. Murty, G. Mainland, I. Rose, A. R. Chowdhury, A. Gosain, J. Bers, and M. Welsh, *Citysense: A vision for an urban-scale wireless networking testbed*, Proceedings of the 2008 IEEE International Conference on Technologies for Homeland Security, pages 583588. IEEE Press, 2008.

[7] M. Wirz, D. Roggen, and G. Troster, *Decentralized detection of group formations from wearable acceleration sensors*, Proceedings of the 2009 IEEE International Conference on Social Computing, page IEEE Press, Aug. 2009.

[8] M. Wirz, D. Roggen, and G. Troster, *A methodology towards the detection of collective behavior patterns by means of body-worn sensors*, Proc. of UbiLarge workshop at Pervasive, 2010.

[9] N. Davies, D. P. Siewiorek, and R. Sukthankar, *Special issue: Activity-based computing*, IEEE Pervasive Computing, 7(2):2021, 2008.

[10] S. Mann, *Humanistic computing: wearcom as a new framework and application for intelligent signal processing*, Proceedings of the IEEE, 86(11):21232151, 1998.

[11] B. Myers, J. Hollan, I. Cruz, S. Bryson, D. Bulterman, T. Catarci, W. Citrin, E. Glinert, J. Grudin, and Y. Ioannidis, *Strategic directions in human-computer interaction*, ACM Computing Surveys, 28(4):794809, 1996.

[12] S. Bosch, M. Marin-Perianu, R.S. Marin-Perianu, J. Scholten and P.J.M. Havinga, *FollowMe! Mobile Team Coordination in Wireless Sensor and Actuator Networks*, Proceedings of the IEEE International Conference on Pervasive Computing and Communications 2009, 9-13 March 2009, Galveston, Texas, USA. pp. 151-161. IEEE Computer Society Press. ISBN 978-1-4244-3304-9

[13] R.S. Marin-Perianu, C. Lombriser, P.J.M. Havinga, J. Scholten and G. Troster, *Tandem: A Context-Aware Method for Spontaneous Clustering of Dynamic Wireless Sensor Nodes*, Proceedings of the First International Conference on Internet of Things (IOT2008), March 2008, Zurich, Switzerland. pp. 341-359. Lecture Notes in Computer Science (4952). Springer Verlag. ISBN 978-3-540-78730-3.

[14] Scholten, J. and Westenberg, R. and Schoemaker, M. , *Trainspotting, a WSN-based train integrity system*, The Eighth International Conference on Networks, ICN 2009, 1-6 March 2009, Gosier, France. pp. 226-231. IEEE Computer Society Press. ISBN 978-0-7695-3552-4.

 ISBN:978-1-61208-113-7

Adaptive Techniques for Elimination of Redundant Handovers in Femtocells

Zdenek Becvar, Pavel Mach
Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Telecommunications Engineering
Prague 16627, Czech Republic
zdenek.becvar@fel.cvut.cz, machp2@fel.cvut.cz

***Abstract*—Dense deployment of femtocells in mobile wireless networks can significantly increase the amount of handover initiations. This paper analyzes new approach to elimination of redundant handovers. The innovative way dynamically updates current value of techniques commonly used for elimination of redundant handovers. The goal is to investigate the efficiency of two handover elimination techniques, i.e., windowing and handover delay timer. Both techniques are modified to enable adaptation of their parameter according to the channel quality related to the users' position in the cell. Furthermore, the impact of proposed modifications on the user's throughput is examined. All simulations are performed in scenario of 4G networks with femtocells. The results show no benefit of adaptive windowing comparing to conventional one. However, the performance improvement is achieved by adaptation of the handover delay timer.**

Keywords-femtocell; handover; Handover Delay Timer, windowing; LTE-A

I. INTRODUCTION

The studies performed in recent year's show that more than 70% of users' traffic is generated from indoors and this ratio is still rising [1]. Deployment of so called femtocells can cope with limited indoor coverage and the cost of the connection in emerging 4G networks. The femtocell is represented by Femto Access Point (FAP) that provides connection of mobile wireless users to a network. The FAP is generally connected to the backbone through a cable connection, xDSL (Digital Subscriber Line) or optical fiber.

The FAP can offer three types of access: close, open, and hybrid. In case of the close access, only a User Equipment (UE) of the FAP's owner (or subscriber) or very small group of users is allowed to enter the FAP. The group of users with access to the FAP is defined by FAP's owner and it is denoted as Close Subscriber Group (CSG). Other users can not access the network via the FAP. Contrariwise, the open access is designed to share full capacity of the FAP by all UEs in its area. The hybrid access combines both open and close accesses. In hybrid access mode, a part of capacity is permanently dedicated to the FAP's owner or to the CSG. The rest of transmission resources can be consumed by other UEs. The open access enables to increase the throughput in specific area by offloading the macro cell [2]. On the other hand, it increases interference in the close area of the FAP.

The deployment of plenty of FAPs can significantly influence the handover decision procedure. The handover is initiated more often since UEs can receive the signal not only from Base Stations (BSs), but also from all FAPs in its neighborhood. In conventional networks without femtocells, the several techniques are defined to eliminate redundant handovers. The most widely used are: Hysteresis Margin (HM), windowing (also known as signal averaging) [3], and Handover Delay Timer (HDT) [4][5], which extends conventional Time-To-Trigger. These techniques can be implemented also in femtocell networks as presented, e.g., in [6][7]. Both papers demonstrate reduction of an amount of the redundant handovers by investigated techniques. However the authors do not investigate a negative impact of techniques on the throughput. In [8], the authors compare the probability of UE's assignment to the FAP that do not provide the best signal quality. The paper shows some tradeoff between a minimum duration of signal averaging and probability of error assignment.

Another approach of elimination of redundant handovers is to adapt the transmission power of FAPs. The proposals of a power control improvement to reduce the number of redundant handovers in femtocells is presented, e.g., in [9][10][11]. All proposals are able to eliminate the redundant handovers. Nevertheless, the advantage of throughput gain due to the open/hybrid access, as illustrated in [2], is also distinctively suppressed.

A modification of HM, which purpose is to eliminate higher ratio of redundant handovers, is defined in [12]. The authors evaluate so called adaptive HM in scenario with macro BSs. The paper assumes precise knowledge of distance between a UE and its serving BS as well as invariant and accurately known radius of macrocells. The radius of all cells is assumed to be the same. Nevertheless, the radius is varying in time and it is neither regular nor symmetric in practice. Moreover, the radius of individual cells is largely different if FAPs are deployed and the exact position of FAPs is not defined by operator as it is in charge of the user. Thus, the cell radius of FAPs cannot be precisely estimated. Therefore technique proposed in [12] cannot be applied into the networks with femtocells. The above mentioned weaknesses are eliminated by considering RSSI

 ISBN:978-1-61208-113-7

(Received Signal Strength Indicator) or CINR (Carrier to Interference plus Noise Ratio) for adaptation of HM value as presented in [13].

The goal of this paper is to investigate the possibility of application of the adaptation into other techniques for handover elimination. The paper investigates impact of the dynamic adaptation of an actual value for windowing and HDT. The simulations performed in this paper are in line with networks according to LTE-A (Long Term Evolution – Advanced) release 10.

The rest of paper is organized as follows. The next section describes the principle of elimination of redundant handovers and its modifications to enable dynamic adaptation. The third section defines simulation scenario and parameters used for evaluation of throughput. The section four contains the results of simulations and their discussion. Last section presents our conclusions and future work plans.

II. ELIMINATION OF REDUNDANT HANDOVERS

A redundant handover (or unnecessary handover) represents a case when the handover is initiated; however it is not completed before a next handover decision is performed. Also the handover frequently repeated between two adjacent cells in short time intervals can be considered as the redundant handover. The redundant handovers are caused by short time channel variation (e.g., fast fading) or by movement of MSs along the edge of the two neighboring cells. As mentioned in previous section, several techniques can be utilized for minimization of the number of redundant handovers. All common methods are based on delaying of the handover execution for some time interval. During this interval, the MS is not connected to the station providing the best quality of communication channel. Therefore, it has negative impact on quality of service offered to the MS due to the utilization of channel with worse quality than a quality of channel available from other BS.

In this paper, two techniques are considered, i.e., windowing and HDT. The third one, HM, was already investigated in [13].

A. Principle of common windowing and HDT

In case of windowing, the handover decision is done if the average value of observed signal parameter (e.g., RSSI, CINR, etc.) from the target BS drops under the average level of the same parameter at the serving BS (see formula (1)). The average value is calculated over a number of samples denoted as Window Size (*WS*).

$$\frac{\sum_{i=1}^{WS} S_i^{Tar}}{WS} > \frac{\sum_{i=1}^{WS} S_i^{Ser}}{WS} \quad (1)$$

where S_i^{Tar} and S_i^{Ser} represent the level of observed signal parameter at the target and serving BS respectively.

The purpose of HDT is to cope especially with temporary drops of a signal level due to fast fading or when a user is located in shadowed places for a short time interval.

Implementation of the HDT is based on the insertion of a short delay between the time when the handover conditions are met and the time when handover initiation is executed. This delay is labeled HDT. The handover conditions have to be fulfilled over the whole duration of HDT to initiate the handover. Generally, the handover is performed if:

$$S_t^{Ser} < S_t^{Tar} \left| t \in (t_{HO}, t_{HO} + HDT) \right. \quad (2)$$

where *HDT* represents the duration of the handover delay timer; and t_{HO} is the time instant when the handover conditions are fulfilled.

B. Adaptive techniques

In the conventional techniques for elimination of redundant handovers, the threshold value (HM, WS, or HDT) is not related to the users' position. Hence, it can be considered as invariant since it is modified by a network only rarely. The adaptive techniques are based on the modification of actual HM value according to the position of the user in the cell. The proposal on adaptive HM is defined in [12]. According to [12], the current HM value is decreasing with the UE's moving closer to the cell boarder as presents the next formula:

$$HM = \max\left\{HM_{\max} \times \left(1 - \frac{d}{R}\right)^4 ; 0\right\} \quad (3)$$

where HM_{max} is the maximum value of HM that can be reached (this value can be set up only in the middle of the cell); d is the distance between the serving BS and the UE; and R is the radius of the serving BS. A modification of adaptive HM is proposed in [13] as the parameters d and R cannot be easily determined neither by the network nor by the UE. This modification considers the signal characteristics (RSSI or CINR) to derivation of current value of HM. The analogical modification should be done for adaptation of WS and HDT. The derivation of actual values for both adaptive techniques is defined by the following equations:

$$WS = max\left\{WS_{max} \times \left(1 - 10^{\frac{CINR_{act} - CINR_{min}}{CINR_{min} - CINR_{max}}}\right)^4 ; 0\right\} \quad (4)$$

$$HDT = max\left\{HDT_{max} \times \left(1 - 10^{\frac{CINR_{act} - CINR_{min}}{CINR_{min} - CINR_{max}}}\right)^4 ; 0\right\} \quad (5)$$

where WS_{max} and HDT_{max} are maximum levels of WS and HDT respectively; $CINR_{act}$ is the actual CINR measured by a UE; $CINR_{min}$ and $CINR_{max}$ are minimum and maximum values in the investigated area respectively.

The $CINR_{act}$ is measured periodically by UEs to monitor the channel state. It is usually performed with purpose of the handover decision. As well as in the case of adaptive HM, the minimum and maximum CINR values have to be determined for the utilization of the adaptive WS and HDT.

 ISBN:978-1-61208-113-7

The $CINR_{min}$ is derived as lowest CINR level at which the UE is still able to receive data. Hence, it is set up to a fix value. Determination of the $CINR_{max}$ is executed via monitoring and reporting of CINR by all UEs connected to the given FAP and than selecting the highest CINR from all known values as the $CINR_{max}$. The exact value of $CINR_{max}$ is permanently updated since the channel conditions are time variant. Therefore, the $CINR_{max}$ is acquired over several samples of CINR measured by UEs. The number of the latest samples utilized for the $CINR_{max}$ derivation is represented by parameter $CINR_{win}$. The optimum value of $CINR_{win}$ is analyzed further in this paper.

III. DESCRIPTION OF SIMULATION PRINCIPLE

A. *Scenario and deployment*

The same scenario and deployment of FAPs and macro BSs as in [13] are considered for the evaluation of both adaptive techniques (see Fig. 1). The scenario contains fifty houses regularly and symmetrically placed along the direct street with length of 500 m. Also all FAPs and BSs are placed symmetrically along the street in the scenario.

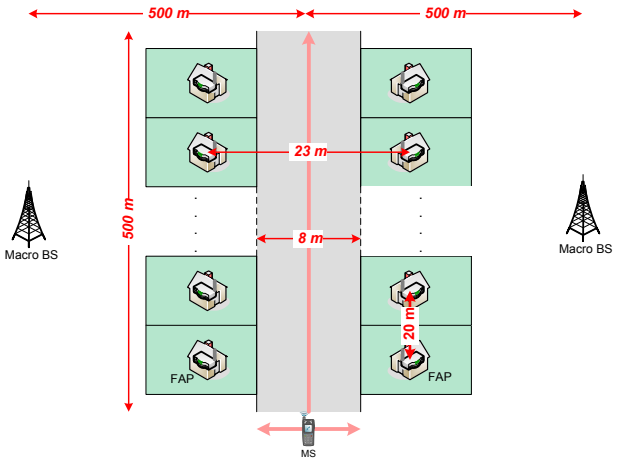


Figure 1. Deployment of FAPs and BSs for simulations.

The users are moving directly along the street with the speed of 1 ms^{-1} until they reach the end of the street. The users are equally distributed over the street width with spacing of 0.2 m. The reporting of measured CINR is executed in periodic intervals of 0.5 s. The signal level received by a UE from a FAP is calculated according to ITU-R P.1238 path loss model for single-storied house. The path loss model includes wall losses and channel variation due to the fast fading and shadowing with standard deviation of 4 dB as defined in [14]. The propagation of BS's signal is in line with Okumura-Hata path loss model for outdoor to outdoor communication [15]. As well, all other simulation parameters, presented in Tab. 1, are set up to be in line with simulations performed by Femto Forum [15].

The amount of handovers is obtained as a number of initiated handovers. It means, if all conditions for the handover initiation are fulfilled, the handover is taken into account no matter if it is finished or not.

TABLE I. SIMULATION SCENARIO AND PARAMETERS

Parameter	Value
Frequency	2 GHz
Channel bandwidth	20 MHz
Transmiting power of BS / FAP	43 / 15 dBm
Height of macro BS / FAP / MS	30 / 1 / 1.5 m
External / Internal Wall Loss	10 / 5 dB
FAP path loss model ITU-R P.1238	20log(f)+28log(d)-24
BS path loss model Okumura-Hata	69.55+26.16log(f)-13.82log(h_B)+(44.9-6.55log(h_B))log(d)-(1.1log(f)-0.7)h_M+(1.56log(f)-0.8)
Channel bandwidth	20 MHz
Noise	-100.97 dBm
Number of simulation drops	25
$CINR_{min}$	-3 dB
$CINR_{win}$	10 ÷ 500

B. *Throughput calculation*

The evaluation of throughput is performed for TDD frame structure of LTE release 10 with uplink–downlink (UL–DL) configuration *"1"* and Special Subframe (SS) configuration *"0"* (see [16] for more details).

In simulations, we assume normal cyclic prefix (seven symbols per subcarrier) and 12 subcarriers per a resource block since those are typical values defined in LTE release 10. The spacing of subcarriers is $\Delta f = 15\ kHz$. The amount of transferred bits depends on Modulation and Coding Scheme (MCS) used for the transmission. The assignment of the MCS is based on signal quality according to Tab. 2 (the values are taken from [17]).

TABLE II. SELECTION OF MCS ACCORDING TO CINR

CINR [dB]	MCS	Transmission efficiency Γ [bits/symbol]
$CINR_{min}$ <CINR <= 1.5	1/3 QPSK	0.66
1.5 < CINR <= 3.8	1/2 QPSK	1
3.8 < CINR <= 5.2	2/3 QPSK	1.33
5.2 < CINR <= 5.9	3/4 QPSK	1.5
5.9 < CINR <= 7.0	4/5 QPSK	1.6
7.0 < CINR <= 10.0	1/2 16QAM	2
10.0 < CINR <= 11.4	2/3 16QAM	2.66
11.4 < CINR <= 12.3	3/4 16QAM	3
12.3 < CINR <= 15.6	4/5 16QAM	3.2
15.6 < CINR <= 17.0	2/3 64QAM	4
17.0 < CINR <= 18.0	3/4 64QAM	4.5
18.0 < CINR	4/5 64QAM	4.8

The throughput of UEs via wireless interface is assumed to be with no limitation caused by the FAP's backbone connection since the FAPs are supposed to be connected to the backbone through a high speed optical fiber.

IV. PERFORMANCE ANALYSIS

The results, obtained by own developed MATLAB simulator, are divided into two subsections according to investigated technique.

A. *Adaptive Window Size*

As it is depicted in Fig. 2, the adaptive WS leads to the significant reduction of performed handovers for low number

 ISBN:978-1-61208-113-7

of averaged samples (roughly up to 7 samples). Then the efficiency of the adaptive technique drops down and the handovers are performed more often. The decreasing efficiency for higher WS is due to the fact that the radius of FAP is very small. Thus, the signal received from the FAP rises and drops rapidly if the user is moving. Therefore, the high WS leads to consideration of samples obtained long time ago with respect to the small FAP radius and users' speed. These samples misrepresent the actual WS and thus the handover is initiated in improper place. Note that the *x* axis in all following figures represents the actual value of WS and HDT for conventional windowing and HDT. In case of WS and HDT with adaptation, the *x* axis expresses WS_{max} and HDT_{max} (see equations (4) and (5)).

The impact of $CINR_{win}$ is only minor for short length of window. The optimum WS_{max} for the adaptive WS is roughly 7 samples since the ratio of performed handovers is the lowest. The efficiency of handover elimination is rising with $CINR_{win}$. However, the results for $CINR_{win}$ equal to 50 and 500 samples are almost the same at WS = 7 samples.

The ratio of eliminated handovers behaves different for conventional windowing with fixed amount of averaged samples. In this case, the amount of initiated handovers is continuously decreasing with growing WS. Nevertheless, the efficiency improvement only by approximately 6% is achieved if WS is increased from 7 to 25 samples. Consequently, Fig. 2 does not proof any benefit in elimination of handovers by implementation of adaptive WS.

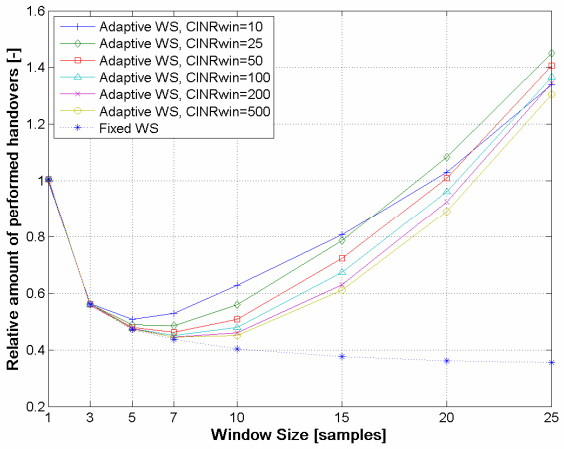


Figure 2. Impact of adaptive WS on the amount of initiated handovers.

Fig. 3 presents the impact of WS on the downlink throughput. This figure shows no considerable difference between adaptive and fixed WS size if WS value is up to 5 samples. Than, the proposed adaptive WS with shorter $CINR_{win}$ is preferable since it leads to the throughput gain.

By combining the results presented in Fig. 2 and Fig. 3 can be observed that the optimum length of $CINR_{win}$ is roughly 50 samples. Both figures further show some throughput gain of adaptive WS. However this gain is at the cost of lower efficiency of handover elimination. Thus the adaptation of WS is not profitable.

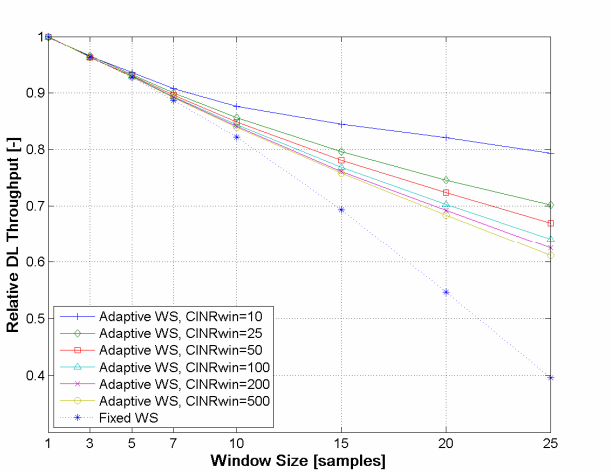


Figure 3. Average DL throughput over WS (for conventional windowing) or WS_{max} (for adaptive WS).

B. *Adaptive Handover Delay Timer*

The impact of HDT adaptation on the amount of handovers and downlink throughput is depicted in Fig. 4 and Fig. 5 respectively. The range of HDT values up to 30 s (*x* axis in Fig. 4 and Fig. 5) can be considered since only slowly moving users (pedestrians) are assumed to perform handover to a FAP. The vehicular users do not spend enough time in the femtocell to complete the handover.

The Fig. 4 shows that the most of handovers is eliminated by HDT of 2 s. Additional prolongation of HDT up to 6 s leads to moderate decrease of the handover amount. The HDT over 6 s does not eliminate any further noticeable portion of handovers. The $CINR_{win}$ influences the results only insignificantly if more than 10 samples is considered.

The conventional as well as adaptive HDT eliminate handovers with the similar efficiency except the HDT = 2 s. For this value, the common HDT outperforms the adaptive one roughly by 5 %. Nevertheless, the efficiency of handover elimination of both adaptive and fixed HDT can be considered as nearly the same for all other values of HDT.

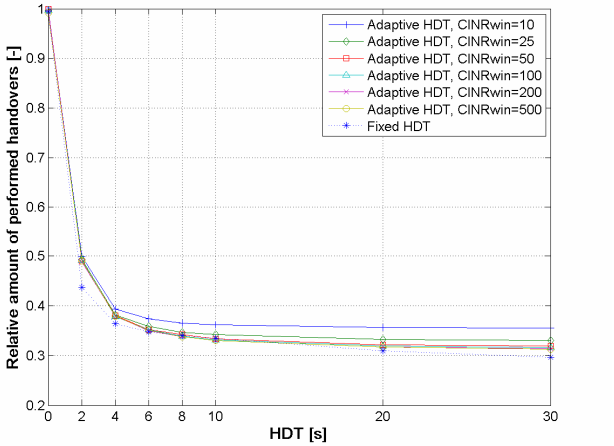


Figure 4. Impact of adaptive HDT on the amount of initiated handovers.

As can be observed from Fig. 5, increasing length of $CINR_{win}$ decreases users' throughput. Hence the shorter length of $CINR_{win}$ is suggested to eliminate throughput drop.

Comparing the fixed and adaptive HDT, significantly more negative impact on the throughput is caused by the technique with no adaptation of current HDT value. The adaptive HDT enables to reach significant gain in the throughput comparing to the conventional one. The gain noticeably rises with HDT duration.

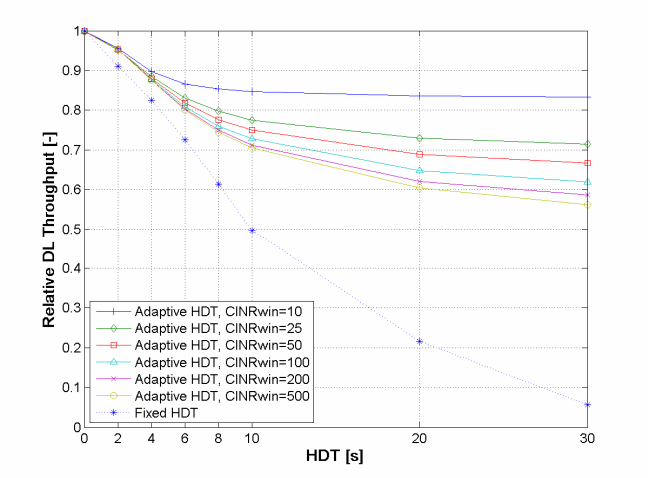


Figure 5. Average DL throughput over HDT (for conventional windowing) or HDT_{max} (for adaptive HDT).

Considering the results presented in Fig. 4 and Fig. 5, the optimum $CINR_{win}$ is roughly 25 samples. This value is same like optimal one for adaptive HM presented in [13]. The most efficient length of HDT is between 4 and 6 s. The adaptive as well as fixed HDT achieves the similar level of handover elimination. Nevertheless, the proposed adaptation of HDT enables throughput gain between 8 % and 13% for the optimal HDT and $CINR_{win}$.

V. CONCLUSIONS

The paper evaluates the efficiency of adaptive WS and HDT for elimination of redundant handovers in the networks with femtocells.

The simulation results show that the adaptation of WS provides the similar results as the windowing technique with the fixed value of WS. On one hand, the adaptive WS leads to some throughput gain, but on the other hand, it also eliminates less handovers. The adaptive duration of HDT can leads to the significant throughput gain while the same elimination of handovers as in case of fixed HDT is reached. The gain is between 8% and 13% for optimal duration of HDT. The optimal length of $CINR_{win}$ is roughly 25 samples.

The future work will be focused on the improvement of handover decision phase in femtocells by considering the handover prediction.

ACKNOWLEDGMENT

This work has been performed in the framework of the FP7 project FREEDOM IST-248891 STP, which is funded by the European Community. The Authors would like to acknowledge the contributions of their colleagues from FREEDOM Consortium (http://www.ict-freedom.eu).

REFERENCES

[1] Informa Telecoms and Media, "Mobile Broadband Access at Home," Aug. 2008.

[2] D. Lopez-Perez, A. Valcarce, G. De La Roche, E. Liu, and J. Zhang, "Access methods to WiMAX femtocells: A downlink system-level case study," Proc. 11th IEEE Singapore International Conference on Communication Systems (ICCS 2008), Nov. 2008, pp. 1657 – 1662, doi: 10.1109/ICCS.2008.4737463.

[3] M. Zonoozi, P. Dassanayake, M. Faulkner, "Optimum Hysteresis Level, Signal Averaging Time and Handover Delay," Proc. 47th IEEE Vehicular Technology Conference (VTC 1997), May 1997, pp. 310 - 313, doi: 10.1109/VETEC.1997.596370.

[4] C. Hoyman, et. al., "Advanced Radio Resource Management Algorithms for Relay-based Networks," Deliverable 2D2 of IST FP6-027675 FIREWORKS project, Jul. 2007.

[5] Z. Becvar, J. Zelenka, "Implementation of Handover Delay Timer into WiMAX," Proc. of 6th Conference on Telecommunications (ConfTele 2007), Peniche, Portugal, May 2007.

[6] M.Z. Chowdhury, W. Ryu, E. Rhee, and Y. M. Jang, "Handover between macrocell and femtocell for UMTS based networks," Proc. 11th International Conference on Advanced Communication Technology(ICACT 2009), Feb. 2009, pp. 237 – 241.

[7] J.-S. Kim and T.-J. Lee, "Handover in UMTS networks with hybrid access femtocells," Proc. 12th International Conference on Advanced Communication Technology (ICACT 2010) , Feb. 2010.

[8] G. Joshi, M. Yavuz, and C. Patel, "Performance analysis of active handoff in CDMA2000 femtocells," Proc. National Conference on Communications (NCC 2010), Jan. 2010, pp. 1 – 5, doi: 10.1109/NCC.2010.5430219.

[9] S. Y. Choi, T.-J. Lee, M. Y. Chung, and H. Choo, "Adaptive Coverage Adjustment for Femtocell Management in a Residential Scenario," Proc. 12th Asia Pacific Network Operation and Management Symposium (APNOMS 2009), Sep. 2009.

[10] H. Claussen, F. Pivit, and L. T. W. Ho, "Self-Optimization of Femtocell Coverage to Minimize the Increase in Core Network Mobility Signalling," Bell Labs Technical Journal, vol. 14, no. 2, 2009, pp. 155 – 184, doi: 10.1002/bltj.20378.

[11] H.-S. Jo, C. Mun, J. Moon, and J.-G. Yook, "Self-optimized Coverage Coordination and Coverage Analysis in Femtocell Networks," Oct. 2009, available online at: http://cdsweb.cern.ch/record/1212431, accessed: Mar. 5, 2010.

[12] S. Lal and D. K. Panwar, "Coverage Analysis of Handoff Algorithm with Adaptive Hysteresis Margin," Proc. 10th International Conference on Information Technology (ICIT 2007), Dec. 2007, pp. 133 – 138, doi: 10.1109/ICIT.2007.68.

[13] Z. Becvar and P. Mach, "Adaptive Hysteresis Margin for Handover in Femtocell Networks," Proc. International Conference on Wireless and Mobile Communications (ICWMC 2010), Sept. 2010.

[14] ITU-R M.2135 Recommendation, "Guidelines for evaluation of radio interface technologies for IMT-Advanced," 2008.

[15] FemtoForum, "Interference Management in UMTS Femtocells," Dec. 2008, available online at: http://www.femtoforum.org/femto/publications.php, accessed: Feb. 23, 2010.

[16] 3GPP Technical Specification 36.300 v 10.0.0, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," Jun. 2010.

[17] C. Yu, W. Xiangming, L. Xinqi, and Z. Wei, "Research on the modulation and coding scheme in LTE TDD wireless network," International Conference on Industrial Mechatronics and Automation (ICIMA 2009), May 2009, pp. 468 – 471.

 ISBN:978-1-61208-113-7

Reconfigurable Tactical Impulse Radio UWB for Communication and Indoor Localization

Thomas Beluch*§, Aubin Lecointre*§, Daniela Dragomirescu*§, Robert Plana*§
*CNRS ; LAAS ; 7 avenue du colonel Roche, F-31077 Toulouse, France
§Université de Toulouse ; UPS, INSA, INP, ISAE ; LAAS ; F-31077 Toulouse, France

***Abstract*—The paper is focused on the design of a new reconfigurable tactical UWB impulse radio enabling indoor communication and localization for military application. The proposed system contains both a physical and a MAC (Medium Access Control) layer. It proposes to answer military needs for information sharing and for indoor localization of infantry. The physical layer is based on IR-UWB (Impulse Radio Ultra WideBand). Its goal is to ensure communication and distance evaluation between the transmitter and the receiver. IR-UWB is necessary for constraints such as low power consumption, reconfigurability, through the wall propagation, fine resolution localization, low probability of detection and interception. These advantages come from the use of a very large bandwidth with very short pulses. The digital baseband implementation of a reconfigurable IR-UWB transceiver is based on a coherent Rake receiver and on a parallel search acquisition. Its synchronization accuracy is of 0.33 ns. This enables distance evaluation with a precision of 10 cm. The MAC layer deals with the multi user access and the positioning. TDMA is used for sharing the channel between users. WiDeCS, a master-slave TOA (Time of Arrival) cross layer mechanism is implemented for time synchronization. This algorithm can also determine the Time of Flight of the IR-UWB signal due to timestamps at the physical layer. This algorithm followed by triangulation and-or angle of arrival techniques leads to 3D positioning.**

***Index Terms*—indoor ; IR-UWB ; localization ; Wireless Sensor Networks**

I. Introduction

Positioning the infantry units is a strategic issue on a battlefield. For many years, Global Positioning System (GPS) has allowed teammates, as well as commanding officers, to know in real time the position of units taking part in a battle. This tactical advantage, as a consequence, is lost in indoor environments such as buildings, or caves. There is a need for indoor positioning equipment for ad-hoc networks. Although millimeter precision is not mandatory for such applications, a position error under 50 cm is sufficient in order to distinguish a teammate and other persons. The main requirements are

- indoor capability and through the wall positioning
- a fast acquisition of teammates' location
- a reliable tracking of moving partners
- a secure transmission channel hard to detect and to jam
- a communication channel for health monitoring

Recent researches have focused on a new way to localize team members in indoor environments. Most of them require the previous installation of reference base stations to locate units in absolutes coordinates.

This work is focusing on the possibility for each unit to obtain the location of every other unit in both direct and indirect range. For that, a system based on Impulse Radio Ultra Wide Band is proposed. The IR-UWB allows an accurate timestamping of emitted and received packets. Over this physical layer, we propose to use a Synchronization and Localization protocol named WiDeCS[1]. WiDeCS uses timestamping at physical level, coupled with two-way ranging to obtain both the time reference offset, and the Time of Flight. The use of Smart antenna allows to obtain an information about the Angle of Arrival, and then determine the position of every node in range by combining ToF information and Angle of Arrival.

In the first section, this paper proposes a possible system using the proposed schemes for localization of moving units. Related works on this subject are then overviewed. The proposed system, and its two major parts - the physical layer, and the MAC layer with synchronization and localization - are detailed in a third part. The last part deals with theoretical performances of the system.

II. Possible integration and operational system assembly

The proposed system is based on the use of two techniques to determine the position of the nodes in range.

A. Smart antenna for fast rough localization

The first step towards a positioning of all the units is to determine roughly the position of neighboring units by the means of Time of Arrival and Angle of Arrival. This step has the main advantage of giving a first rough estimation to the user.

B. Map generation

In the same time, an array of distances and directions is computed with the help of WiDeCS and is stored inside the MAC layer. The MAC layer shares the information of distance between nodes with nodes in range. A map can then be determined owing to these information and a trilateration algorithm.

III. Related Work

Indoor positioning is a challenging subject gathering resources from all around the world. Rough localization has been made possible in the case of widely used wireless networks

 ISBN:978-1-61208-113-7

such as WiFi and Bluetooth. However, the achievable accuracy is limited by a major drawback of these technologies, the use of a narrow band, making event time stamping hard to improve [2].

Recent advances on this subject have improved the spatial resolution of positioning systems. Most of the proposed systems use Ultra Wide Band (UWB) transceivers. The UWB technologies allow large band emission, which present fast changes in the time domain. These events can be used for measuring the Time of Arrival (ToA) or Time Differential of Arrival (TDoA). UWB makes it possible to implement low power and low complexity systems. The three most widely spread methods for UWB localization are [3]:

- Triangulation uses one distance and 2 angle informations to compute the position of a node when the 2 reference base stations have known positions. Its advantages are the small number of base stations necessary to determine the position coupled with no necessary synchronization between base stations. However, a high accuracy of the measured Angle of Arrival (AoA) is necessary.
- Trilateration is a method similar to triangulation with the difference that it uses a ratio between distances measured relatively to at least three base stations. This distance can be estimated from the ToA. However, a high synchronization is required between all the stations.
- Multilateration is derived from trilateration. Instead of using absolute ToA to determine the distance, the TDoA allows to not synchronize the localized node with the base stations. Those base stations, on the other side, still need to be synchronized.

A rough estimation of the relative position can be determined if an information of distance is coupled with the Angle of Arrival. The Received Signal Strength (RSS)[4] helps determining the approximate distance owing to the knowledge of the channel path loss. The other possible way to know the distance is based on a ToA measurement. However, this technique's performances are linked to the synchronization performances. UWB appears to be a solution to the ToA measurement issue. For example, Clarke et al. compared commercially available positioning systems based on different modulation (e.g. Wi-Fi, signal strength, radio frequency (RF), ultrasound, and UWB) [2]. Implementations based on IR-UWB have been proposed in the literature. Rahmatollahi et al. [5] described a solution using base stations coupled with IR-UWB. However, this system cannot be used without a previous deployment. The main challenge in such systems is to detect the shortest path among all the multipath pulses. Indeed, the most important pulse is not always the one carrying the greatest amount of energy. Recent publication show propositions based on the detection of the first pulse to arrive in the receiver [6] The proposed system embeds the localization protocol in every node in order to localize the nodes relatively to each other instead of positioning them relatively to a previously deployed infrastructure. It is then possible to localize persons equipped with such system even in the case of a first visit of the indoor environment.

A. System view

The localization system is based on a network communication stack. Each node must gather Time of Flight (ToF) information regarding its data exchanges with the other nodes and compile them in an array. This array is the utilized at application layer to determine location of all the nodes. The figure 1 shows an example of deployed system.

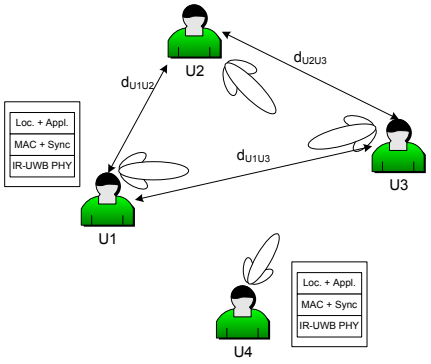


Fig. 1. Operational use case

A smart antenna is placed on the helmet to cover all the possible incoming directions. These antennas are connected to multiple instances of the physical layer. The first instance detecting the pulses and decoding them is supposed to be attached to the antenna pointed in the direction of the shortest path between the transmitting node and the receiving node. This shortest path is assumed to be the Line Of Sight path owing to the through the wall propagation performances of IR-UWB.

The physical layer used in this system is an Impulse Radio Ultra Wide Band protocol. This protocol is based on the use of short pulses spread over an ultra wide bandwidth. The possible achievable synchronization is then directly linked to the occupied bandwidth and the receiver resolution. Details about the use of IR-UWB in such systems are given in a first time.

Once the physical layers are synchronized, it is then important to synchronize the clocks of the discussing nodes in order to determine the Time of Flight (ToF) as precisely as possible. The WiDeCS synchronization scheme is used for this purpose [1]. The MAC layer also computes the ToF calculations and delivers them to the application dedicated to maintaining a map of neighboring teammates.

B. Physical layer specificities

The proposed physical layer is based on IR-UWB (Impulse Radio Ultra WideBand). Its goal is to ensure communication and distance evaluation between the transmitter and the receiver. IR-UWB uses pulse modulation for transmitting information. IR-UWB suits to constraints such as low power consumption, reconfigurability, through the wall propagation,

 ISBN:978-1-61208-113-7

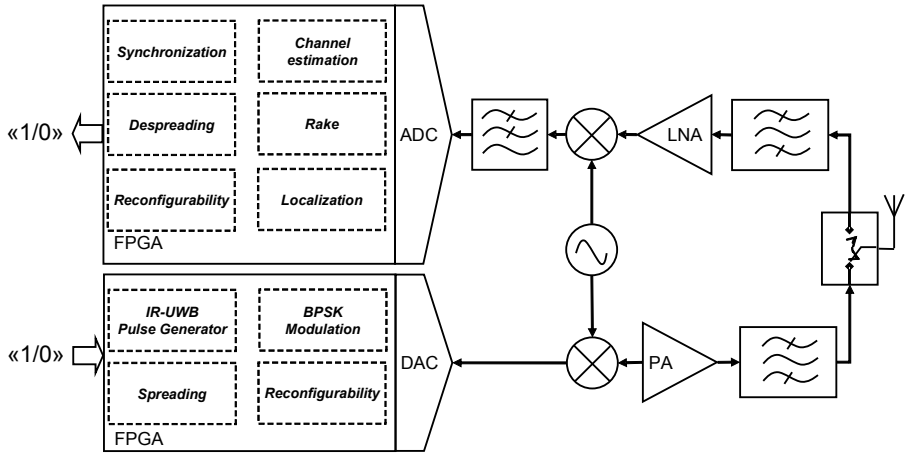


Fig. 2. Proposed IR-UWB transceiver implementation with localization capability.

fine accuracy localization, multi users scalability, low probability of detection and interception [7]. In addition, IR-UWB technique has a low probability of pulse collision enabling the support of multi user scenario with low complexity MAC-level mechanism. The use of very short impulse confers an excellent behavior regarding the fading caused by multipath in comparison with the classical narrow band techniques [8]. IR-UWB is a very promising technology for both low power communication and localization thanks to the use of a very large bandwidth with very short pulses[9]. The localization and the synchronization accuracy directly depend on the occupied bandwidth of the IR-UWB signal. This latter defines the resolution of the receiver, which is equal to the inverse of the occupied bandwidth. The finer it is, the more precise the synchronization and the localization are. Time of arrival (ToA) localization techniques suits to IR-UWB features [9]. The shorter the IR-UWB pulse is, and the better the localization accuracy is. The goal of ToA is to evaluate the time of flight of the pulse between the transmitter and the receiver, by measuring the time of arrival of the received pulse. Thus, the proposed system is able to evaluate the distance between the transmitter and the receiver. At this stage it is only a 1-D positioning. For achieving fine localization accuracy a fine estimation of the pulse time of arrival is needed, implying an efficient synchronization algorithm at the receiver. This performance requirement is also need for achieving IR-UWB communication. The localization is a classic IR-UWB communication. The better the IR-UWB communication performances are, the better the IR-UWB localization performances are.

The UWB occupied bandwidth is large and the UWB channel has a dense multipath behavior, these confer to the synchronization task a high level of complexity [10][11]. The receiver resolution is small and equal to the inverse of the occupied bandwidth. The number of multipath components resolved is then large. The short impulse duration, the large space search and the low power UWB signal levels are implied in the high complexity of the IR-UWB synchronization. An efficient synchronization technique has to be fast and precise for, respectively, avoiding a large decrease of the signal to noise ratio and decreasing its cost. The synchronization criteria can be used for evaluating the localization technique. In IR-UWB many multipath components can be seen as solution of the synchronization even if they result from a NLOS contribution. This can imply an error in the distance estimation mechanism since a NLOS multipath is not directly linked to the distance between the transmitter and the receiver.

The proposed IR-UWB receiver, assuring both communication and distance evaluation, is based on a coherent BPSK Rake receiver and on a parallel search synchronization. The proposed IR-UWB transceiver allows fine localization at low cost and low power. It achieves reconfigurable performances in data rate, bit error rate, radio range, spectrum occupation, power consumption, synchronization accuracy, processing gain, transmitted power and pulse duration [12]. This reconfigurable behavior allows to efficiently fulfill the distinct applications needs and the evolutions of the environment at the best cost. [12] demonstrates that decreasing the synchronization accuracy implies a decrease of the power consumption. Thus in function of the localization accuracy requirement, defined by the application needs, the proposed reconfigurable IR-UWB transceiver can adapt its power consumption for increasing its life operating duration.

The synchronization technique used allows to determine the first multipath received. The proposed implementation of the IR-UWB transceiver is described in fig. 2. It is implemented according to the mostly digital implementation thus the performances directly depend on the digital to analog and analog to digital converters (DAC and ADC).

By considering the ADC sampling frequency Fe, the achievable synchronization accuracy ϵ, in seconds, of the receiver is defined by:

$$\epsilon[s] = \frac{1}{F_s} \tag{1}$$

The localization accuracy β, in meters, is obtained as follow:

$$\beta[m] = v_{UWB} \times \epsilon \tag{2}$$

 ISBN:978-1-61208-113-7

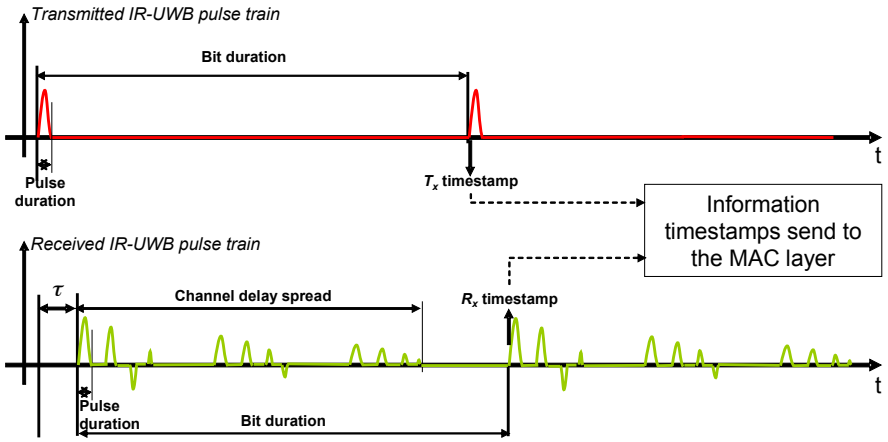


Fig. 3. First multipath component timestamp

where vUWB is the propagation speed of the IR-UWB in the considered environment. The proposed receiver is implemented on FPGA Virtex 4 board with a 3 GSPS time interleaved ADC. In this case, the synchronization accuracy is equal to $\epsilon = 0.33ns$. This enables distance evaluation with a precision of $\beta = 10cm$. The information conveyed by the proposed PHY to the MAC layer allows only 1-D positioning, i.e. distance estimation between the transmitter and the receiver as described by the figure 2. For 2-D or

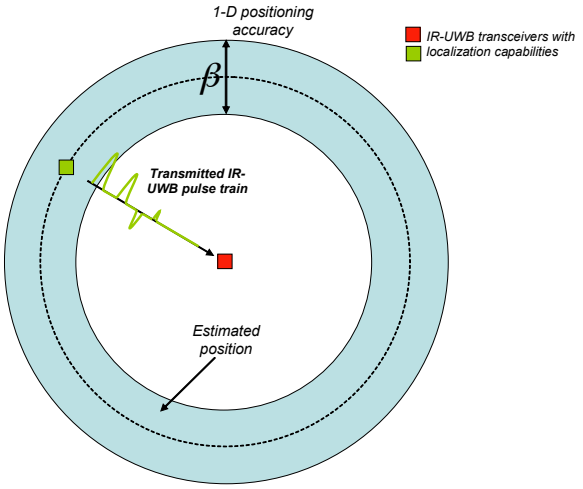


Fig. 4. Illustration of the 1-D positioning offered by IR-UWB transceivers.

3-D positioning high level algorithm must be used such as triangulation, trilateration, or multilateration. The PHY layer sends to the MAC layer two timestamps information. At the transmitter side, the PHY layer warns the MAC layer when the first pulse containing information is sent. Whereas at the receiver side, the PHY layer wars the MAC layer when it detects the first multipath component of the first received pulse of information. Even in case of NLOS communication the use of the first multipath component is required for reducing the average distance estimation accuracy as illustrated in the figure 3.

C. MAC Layer and cross layering for distance evaluation

The MAC layer used for this system is based on a static TDMA frame with pre-assigned slots which includes an implementation of the Wireless Deterministic Clock Synchronization (WiDeCS) scheme. WiDeCS is a protocol allowing high performance synchronization in wireless sensor networks while keeping a low power consumption.

The proposed synchronization protocol takes advantage of the restrictions linked to the application to make strong assumptions and simplify the development of both the MAC layer and the synchronization protocol.

1) General description: The Wireless Deterministic Clock Synchronization (WiDeCS) protocol is based on the planning of transmissions, and the respect of this planning. It is designed for propagating the master time reference of a star network to all the slaves.

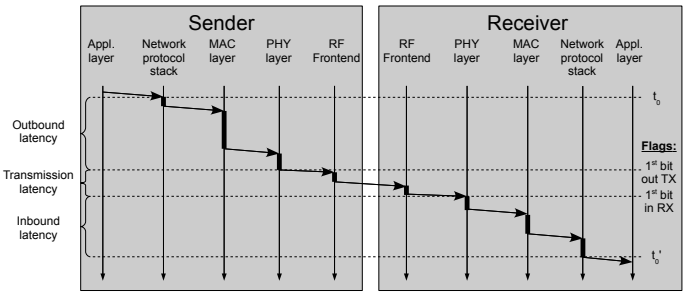


Fig. 5. Latencies involved and flag positioning for WiDeCS

For first experiments, a Time Division Multiple Access MAC layer is used. This MAC layer divides time in slots attributed to the nodes. The figure 5 details sources of delays in the network link. Delays are measured on the first effective bit (events linked to preambles and serialization have to be de-embeded). Efforts have been made to place this time stamping as close as possible to the effective channel, thus reducing uncertainties on the characterization of the propagation time on the channel. WiDeCS uses time stamping at the time of the first effective bit at the output of the PHY layer in emission, and the

 ISBN:978-1-61208-113-7

first effective bit in input of the PHY layer in reception. These flags then enable precise measurement of the propagation time except for the jitter linked to RF front-ends. We name these flags TX_ONGOING for the first bit of effective data out of the PHY transmitter, and RX_ONGOING for the first bit of effective data entering the PHY receiver.

Owing to this MAC layer, each node of the network is supposed to talk at precise moments. Measuring delays between expected receiving times and actual ones helps determining the clock offset and the propagation time. WiDeCS Synchronization protocol uses possibilities relative to time division to determine the clock offset between each node and the master. The figure 6 shows the different informations gathered on different sides, and sent to the slaves by the master. The

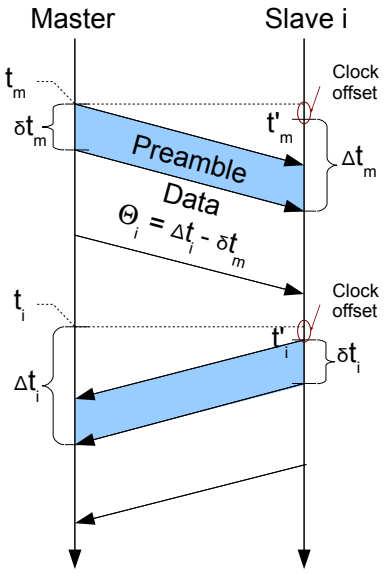


Fig. 6. Exchange of timing information

complete synchronization is based on 2 steps.

Phase 1: Pre-synchronization

This step is performed by every slave when joining the network.

The master has the same communicating hardware the slaves have. The function of the master is to create a frame, and regularly send a preamble of the TDMA frame. The slaves, when reset, or desynchronized, switch to reception only mode, and wait for the master node to indicate the beginning of the frame. When this message is detected, the slave resets its *frame counter* to a predefined reset value in order to begin transmitting data in the dedicated slot. When this step is passed, the slave switches to phase 2 of the synchronization process.

Phase 2: Fine-synchronization

Once pre-synchronized, the nodes have a clock offset of 1 μs compared to the master clock.

This step consists in determining the offset with the greatest precision possible. When the slave is allowed to transmit data, it does it at one precise moment t_i of the slot. This date is coded in the protocol and is known by every slave, and more importantly by the master. Data is sent to the PHY layer when $t = t_i$ and a capture of the timer is done when the TX_ONGOING flag is set. t_i is then subtracted to this date, and the result is stored as δt_i. Note that small deltas (δ) are used for delays in transmitted packets, whereas capital deltas (Δ) are for delays in received packets The following equations can be deducted from fig 6. The equation 3 represents the clock offset correction applied to the slave node in WiDeCS scheme.

$$\Delta t_{clk} = \frac{(\Delta t_i + \delta t_m) - (\delta t_i + \Delta t_m)}{2} \quad (3)$$

In the case of distance evaluation, the Time of Flight is necessary to determine the distance between the nodes. The variant equation 4 deducts this Time of Flight information from the same captured deltas.

$$ToF = \frac{(\Delta t_i - \delta t_m) + (\Delta t_i - \delta t_m)}{2} \quad (4)$$

WiDeCS can be used in every frame to maintain a clock as precisely synchronized as possible for all the length of the measurements. Another possibility is to train the slave with an optimal number of frames, and then turn the reception module in sleep mode for as long as blindness is allowed in the network.

Phase 3: Distance evaluation sharing As explained in phase 2, WiDeCS scheme does estimate the distance only between the paster and every slave. A modification to this scheme is then necessary to allow distance evaluation between every node. This modification consists in measuring every delay between planed reception and effective reception of every node's packet. An array of such delays cans the be integrated in every node's packet in order to calculate the ToF between all the nodes.

The equation 5 shows the information transmitted between slaves

$$\Theta_{ij} = \Delta t_j - \delta t_i \quad (5)$$

where i is the node transmitting the Θ_{ij}, and j the node whose delays are spoken of.

2) Application specific improvements and simplification: Usually, a network is composed of nodes coming and leaving the network. The number of nodes changes with time. In the context described before, the network is planed and parametrized before its deployment. Simplifications are then allowed, without risking uncontrolled behavior. The simplifications used in this paper are listed below, and reasons are given for their existence.

- Fixed number of nodes per piconet and pre-programmed time slots:
  As we explained before, the network architecture is fixed, and planned before deployment of the team. The number of nodes for each network is then decided upon the number of persons to localize.
- No jamming between neighboring networks:
  IR-UWB has a low pulse collision probability. The use of directive smart antenna - implying Spatial Division

 ISBN:978-1-61208-113-7

Multiple Access (SDMA) - reinforces this fact. This modulation then allows multiple TDMA networks to exist on the same frequency band. This is possible by the means of a tradeoff between number of networks, and each network's data rate.

- Symmetrical links:
  Every node of the network is fabricated using the same hardware. Delays in RF front-ends are then similar from one node to another.

IV. Proposed system's performances

The IR-UWB modulation is based on the emission of short pulses covering an Ultra Wide frequency Bandwidth. These short pulses are then a very precise means of getting the information of when it left the transmitter, and when it entered the receiver. Considering that the physical layer described previously allows a synchronization with the first received pulse. The resulting first pulse detected is then the one that followed the shortest possible way from the transmitter to the receiver. The proposed PHY layer uses a synchronization technique not based on the acquisition of the highest energy multipath component. This approach reduces the error due to NLOS scenario. The proposed PHY layer in synchronized on the first viable multipath component assuring communication.

The proposed transceiver achieves reconfigurability in data rate (from 4 to $125Mbits/s$), chip duration (from 8 to $32ns$), pulse duration (from 1 to $4ns$), bit duration (from 8 to $256ns$), processing gain (from 1 to 8 pulses/bit), occupied bandwidth, radio range, BER performance, synchronization accuracy (from 0.33 to $2.64ns$), transmitted power density, duty cycle (from 3 to 50 $\%$), power consumption, spectrum occupation, and maximum supported UWB channel delay spread (from 4 to $31ns$) [12]. These performances depend on the used FPGA and DAC/ADC. In our implementation, a FPGA Virtex 4 is used and DAC/ADC at $3GSPS$. Reconfigurability allows to support a large range of operating scenarion with distincts requirements.

WiDeCS scheme [1], on the other hand has reached synchronization accuracies of 1 clock period in simulations, and of 3 clock periods on a specific testbench. The deductible accuracy for the synchronization is then $\Delta t = 0.33ns \times 3 = 1ns$. The distance evaluation accuracy is then of :
$\Delta d = c \times \Delta t = 30cm$ in vacuum.

The Smart antenna adds an information of Angle of Arrival. However, directional antennas often have radiation patterns with main lobe's angle above $15°$. The angle error is then $\pm 7.5°$.

The localization performances are linked to the communication performances. The proposed system achieves a 30 cm resolution 3-D positioning when using triangulation.

V. Conclusion

This paper proposed a physical and a MAC layer adapted to localization of nodes without pre-deployed spatial references. The estimated distances are measured for every node and can be tranfered through the data link for each node to calculate a map of partners' positions. As an opening, a possible system integration including the use of Smart Antennas for rough localisation of neighboring nodes is proposed.

Acknowledgment

The authors acknowledge the French Defense Agency (DGA) for funding the doctoral studies of Thomas Beluch and Aubin Lecointre.

References

[1] T. Beluch, D. Dragomirescu, F. Perget, and R. plana, "Cross layered synchronization protocol for wireless sensor networks," *Networks, 2010. ICN '10. Ninth International Conference on*, pp. 167–172, 2010.

[2] D. Clarke and M. Park, "Active-rfid system accuracy and its implications for clinical applications," *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*, pp. 21 – 26, 2006.

[3] Y. Huang, Y. Lu, H. Chattha, X. Zhu, I. Hewitt, and S. Hussain, "Uwb antennas for radio positioning systems," *Antennas and Propagation, 2009. EuCAP 2009. 3rd European Conference on*, pp. 3779 – 3782, 2009.

[4] J. Ryoo, H. Choi, and H. Kim;, "Sequential monte carlo filtering for location estimation in indoor wireless environments," *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE*, pp. 1 – 2, 2010.

[5] G. Rahmatollahi, M. Guirao, S. Galler, and T. Kaiser, "Position estimation in ir-uwb autonomous wireless sensor networks," *Positioning, Navigation and Communication, 2008. WPNC 2008. 5th Workshop on*, pp. 259 – 263, 2008.

[6] M. Kuhn, J. Turnmire, M. Mahfouz, and A. Fathy, "Adaptive leading-edge detection in uwb indoor localization," *Radio and Wireless Symposium (RWS), 2010 IEEE*, pp. 268 – 271, 2010.

[7] G. Aiello and G. Rogerson, "Ultra-wideband wireless systems," *Microwave Magazine, IEEE*, vol. 4, no. 2, pp. 36 – 47, 2003.

[8] M. Win and R. Scholtz, "On the robustness of ultra-wide bandwidth signals in dense multipath environments," *Communications Letters, IEEE*, vol. 2, no. 2, pp. 51 – 53, 1998.

[9] S. Gezici, Z. Tian, G. Giannakis, H. Kobayashi, A. Molisch, H. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks," *Signal Processing Magazine, IEEE*, vol. 22, no. 4, pp. 70 – 84, 2005.

[10] S. Aedudodla, S. Vijayakumaran, and T. Wong, "Timing acquisition in ultra-wideband communication systems," *Vehicular Technology, IEEE Transactions on*, vol. 54, no. 5, pp. 1570 – 1583, 2005.

[11] J. Ibrahim and R. Buehrer, "Two-stage acquisition for uwb in dense multipath," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 4, pp. 801 – 807, 2006.

[12] A. Lecointre, D. Dragomirescu, and R. Plana, "Largely reconfigurable impulse radio uwb transceiver," *Electronics Letters*, vol. 46, no. 6, pp. 453 – 455, 2010.

 ISBN:978-1-61208-113-7

CQI Reporting Imperfections and their Consequences in LTE Networks

Kari Aho, Olli Alanen
Magister Solutions Ltd.
Rautpohjankatu 8
FIN-40700 Jyväskylä, Finland
firstname.lastname@magister.fi

Jorma Kaikkonen
Nokia
Elektroniikkatie 3
FIN-90570 Oulu, Finland
firstname.lastname@nokia.com

Abstract — **In modern wireless networks, the signal quality in wireless channel is estimated based on the channel quality measurements. The measurement results are used to select appropriate modulation and coding scheme for each transmission. The target of the link adaptation is to reach the desired block error rate operation point. Operation point and system performance could potentially be compromised by non-consistent / biased channel quality indicator reporting caused by, e.g., differently calibrated user equipments or hardware inaccuracies. This paper evaluates the extent of that phenomenon through different combinations of traffic types, bias settings and system loads by the means of fully dynamic system simulations. The in-depth results verified that on the system level the performance is not significantly impacted by reporting imperfections. Long term evolution is used as an example technology in this study, but the same concepts are applicable to other wireless technologies also.**

Keywords: Link adaptation, BLER target

I. INTRODUCTION

The tremendous success of wireless cellular *High Speed Packet Access (HSPA)* networks together with a "push" from competing technologies has fueled the development of cellular technologies even further. *Third Generation Partnership Project (3GPP)* has already completed first specification release of *Long Term Evolution (LTE)* [1] [2] which is considered to be the successor of HSPA.

LTE utilizes more simplified architecture and new radio access technologies, namely *Orthogonal Frequency Division Multiple Access (OFDMA)* in the downlink and *Single Carrier Frequency Multiple Access (SC-FDMA)* in the uplink. By introducing these changes among others the 3GPP has set a range of strict performance requirements for LTE [3]. For instance, LTE should achieve 2-4 times higher spectral efficiency than Release 6 HSPA is capable for.

When compared to the HSDPA, the adaptation to fast wireless channel variations LTE utilizes different techniques, since the transmission power is constant in the downlink. First of all, the *Modulation and Coding Scheme (MCS)* is adapted with frequent interval to the channel quality, based on the *User Equipment (UE)* feedbacks. Secondly, the *evolved NodeB (e-NodeB)* has capability to perform *Frequency Domain Packet Scheduling (FDPS)* to allocate the most suitable resources for the UEs. The purpose of the *Link Adaptation (LA)* is to handle the feedback information gotten from the UEs and then perform the selection of the appropriate MCS for the UE based also on the information about the allocation position in the frequency domain.

Channel Quality Indicator (CQI) plays a key role in the link adaptation process. It is a message sent by UE to e-NodeB describing the current downlink channel quality of the UE. It is measured from the reference symbols transmitted by e-NodeBs. The CQI measurement interval, measurement resolution in frequency domain, reporting mechanisms, etc. are all configurable parameters. These parameters have a tremendous impact on the system performance and their performance is studied e.g. in [4][5].

The *Inner Loop Link Adaptation (ILLA)* has the first hand responsibility on selecting the suitable MCS for the UE. The selection is done based on the mapping between the measured *Signal to Interference plus Noise Ratio (SINR)* of the reference symbols to the most appropriate MCS for an allocation. For various reasons the ILLA does not however always provide the perfect adaptation and therefore Outer Loop Link Adaptation (OLLA) function is also needed. The target of the OLLA is to adapt the MCS selection to provide certain *Block Error Rate (BLER)*. The target BLER (s.c. Operation Point) is usually set to provide optimal performance depending on whether retransmission mechanisms like *Automatic Retransmission reQuest (ARQ)* and/or *Hybrid ARQ (HARQ)* are utilized.

II. RESEARCH PROBLEM AND MOTIVATION

LTE UE radio transmission and reception requirements are specified in [6]. One of these requirements is related to how tightly the BLER operation point should be set. Operation point and system performance could potentially be affected by non-consistent CQI reporting by the UE. In other words non-consistent CQI reporting could lead to having suboptimal BLER operation point. Non-consistent CQI reporting by the UE can be caused by e.g. hardware inaccuracy, misconfiguration or calibration.

The purpose of this study is to evaluate how system level performance is affected if UEs report biased (more aggressive and/or non-aggressive) CQI values instead of the ones that they actually should in their current radio channel conditions. Thus,

 ISBN:978-1-61208-113-7

bias is directly related to the initial offset value of *Outer Loop Link Adaptation (OLLA)* which is then being corrected. In this study bias is referred also as initial LA/OLLA offset.

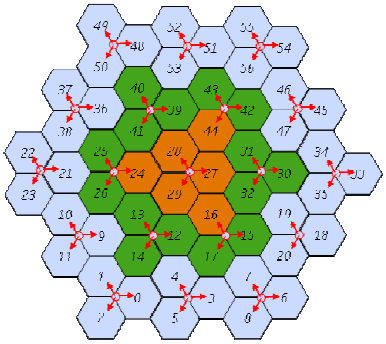


Figure 1. Simulation scenario

III. SIMULATION METHODOLOGY AND ASSUMPTIONS

This study has been performed using a fully dynamic time driven system simulator which supports both uplink and downlink directions with a symbol resolution. In this study only downlink direction is simulated in detail and uplink traffic is considered as ideal to keep the scope of this study within reasonable limits.

We have used periodically reported full CQI information (separately from each *Physical Resource Block (PRB)*)) in these studies to show the impact of bias in worst-case scenario as very accurate CQI information would be available without the bias. The actual bias is studied with two different alternatives: fixed and random bias. Fixed bias means a situation where all terminals have (the same) fixed bias in the beginning of the call whereas with random bias the terminals have bias set according to uniform distribution. Both in the random and fixed bias cases, the bias is constant during the whole call for each UE.

As implied earlier, in this study we assume both an inner loop and an outer loop LA unit. The OLLA algorithm imposes an offset margin that is subtracted from the SINR measurements in the CQI manager before being used by the inner loop LA to estimate the supported data rate, and modulation and coding scheme. The OLLA algorithm aims to control the experienced average BLER for the first transmissions, and it follows the same principle as the traditional outer loop power control algorithm for dedicated channels in IS-95 and WCDMA and for HSDPA. Hence, if an A*cknowledgement (Ack)* is received for a first transmission, the offset factor, A, is increased by A_{up} decibels (defined with a parameter), while it is decreased by A_{down} decibels if a *Negative Acknowledgement (Nack)* is received. Offset factor has limit for maximum and minimum to prevent situations where channel conditions change significantly and OLLA would take very long time to shift the offset back to the other way. The ratio between the step up and down determines the average BLER that the OLLA converges to, i.e.

$$BLER = 1 / (1 + A_{up} / A_{down}). \quad (1)$$

Simulations have been conducted with *Constant Bit Rate (CBR)* type of service which has certain amount of source data and thus certain amount of packets (varied throughout the simulations). See more detailed parameters from TABLE III. Different file sizes do not model any specific applications but are rather just selected to see at which size the CQI bias does have an impact. The smallest sizes like 10kB and 50kB are, however, very small when considering the traffic volume of the modern network applications.

These studies have been conducted in a macro cellular scenario presented in Figure 1. The scenario consists of 19 base stations where two inner tiers (i.e. the orange and green areas) are the one were mobiles are allowed to move. Statistics are collected from the innermost tier (orange cells). Third tier (i.e. cells indicated with light blue colour) are normal active BSs which have background load adapting to statistic BS load. In addition to the adaptive load of the third tier, in this study also the two innermost tiers are adjusted to have minimum level of cell load (0-100 %) which is reached generating artificial (background) load if UEs (avg. 10 per cell) themselves do not reach the target.

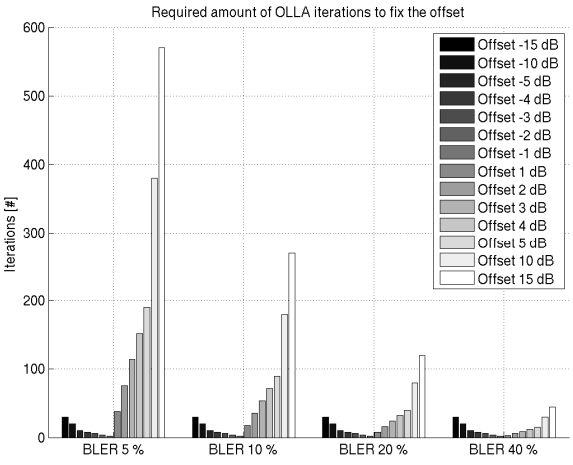


Figure 2. Example of required OLLA iterations to meet the BLER target

IV. PREANALYSIS

As described above, the biased CQI reporting impacts to the performance of OLLA as it needs to fix also the bias in addition to its normal operation. The purpose of this section is to briefly analyze the background of OLLA operation and how the performance would change due to the bias.

TABLE I. EXAMPLE OLLA PARAMETERS

BLER	A_{up}	A_{down}
5%	0.5	~ 0.026
10%	0.5	~ 0.056
20%	0.5	0.125
40%	0.5	~ 0.333

 ISBN:978-1-61208-113-7

Normally, when the UE sends its CQI reports to the e-NodeB the scheduler will do the scheduling assignments according to the received and processed information. The purpose of the e-NodeB link adaptation units is to modify the received CQI information and thus allocated MCS so that certain BLER target is met. This means in practice that UE will experience delay before the certain level of BLER is converged. The delay depends on how high difference there is between OLLA starting point and the actual channel situation. Figure 2. illustrates an example of how many OLLA iterations it takes to fix the offset between the starting point of OLLA and the actual channel situation of the UE. The numbers of iterations are calculated for different BLER levels according to the Equation (1) with A_{up} being 0.5 dB. The resulting Adown values for each of the used BLER percentages can be found in TABLE I. As the figure shows, positive offsets, i.e., situations where the OLLA starts with more conservative MCSs are changed with relatively low pace when compared to the negative offsets where MCSs would be more aggressive. The reasoning behind this is simple, too aggressive MCSs can cause BLER levels to rise rapidly and thus affect to quality of service that users experience whereas more moderate MCSs are likely only to drop the BLER levels. The downside of more moderate MCSs is, however, the drop in user throughput levels, presuming that packets would go through with more aggressive MCSs.

When considering that the bias could further increase the offsets it is possible that system level performance would be affected. However, when considering that the range of actual bias values that would present in the network should remain on quite low level (+-1 dB) the impact should be able to be mitigated quite well by the OLLA unit.

TABLE II. MAIN SIMULATION ASSUMPTIONS

Feature/Parameter	Value / Description
Operational Bandwidth	10 MHz
Duplexing	FDD
Number of sub-carriers	600
Network synchronicity	Asynchroneous
Sub-frame length	1 ms
Cell layout	57 hexagonal macro cells
NodeB Inter site distance (ISD)	500 m
Multipath channel	Typical Urban
UE velocity	3 kmph
UE receiver	2 Rx MRC
Outerloop link adaptation	BLER target 0.2 A_{up} 0.5 dB A_{down} 0.125 dB Max offset 15 dB Min offset -15 dB
Channel quality indicator	Measument period 5 ms PRBs per CQI 6 Reporting delay 2ms SINR error variance 1 dB Bias +-[0, 1, 2, 4]

TABLE III. CBR TRAFFIC ASSUMPTIONS

Feature/Parameter	Value / Description
File size	[10, 50, 100, 200, 1000, 5000] kbytes
Packet size	1500 bytes
Packet inter-arrival time	1 step

V. SIMULATION RESULTS

The system level performance is evaluated in this study mainly through normalized *Spectral Efficiency (SE),* user throughput, first transmission BLER per call and distribution of OLLA offset collected at the end of the call. SEs and user throughputs are normalized so that bias 0 dB, i.e., no bias case is the reference point. User throughputs are presented as percentile bars and e.g. 10 percentile bar height means that 10 percent of the calls experience throughput of that or less.

A. Performance with fixed bias

Spectral efficiency of CBR type of service with different source data amounts is illustrated in Figure 3. As that figure shows the impact of fixed bias is moderately sensitive to the amount of data that is transmitted during the call. If the calls are very short (deductable from the user throughputs and the amount of source data, illustrated in Figure 4. and Figure 5.) the performance can be impacted in terms of SE, depending on the magnitude of the bias. The impact is higher if more aggressive bias (negative values) i.e. higher MCS is selected than the actual CQI would imply. However, if the call length is more realistic, i.e. there is more data and the packets during the call; the performance starts to become more balanced.

User throughputs for 10 and 90 percentiles illustrated in Figure 4. and Figure 5. show that the trend is similar to SE figures above, i.e., the impact of bias starts to diminish once there is reasonable amount packets/data during the call. Moreover, small and moderate positive bias (non-aggressive) can even improve the performance in terms of 90 % user throughput where the MCSs are generally quite high. On another hand more aggressive bias results in 10-15% loss for UEs in similar situation.

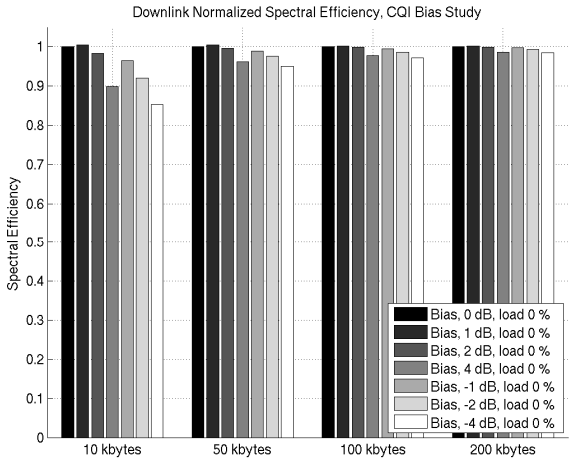


Figure 3. Normalized Spectral Efficiency, Fixed Bias

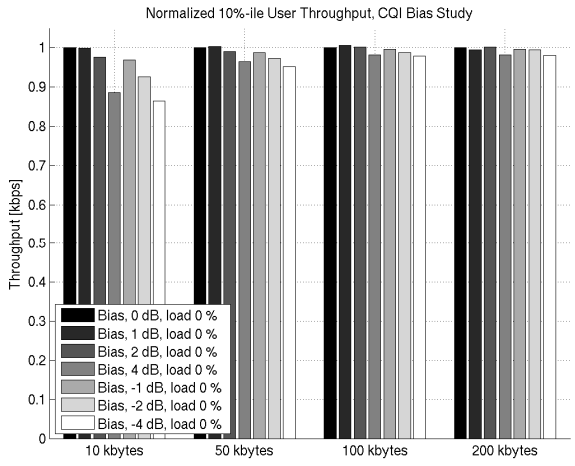


Figure 4. User Throughput, 10 Percentile, Fixed Bias

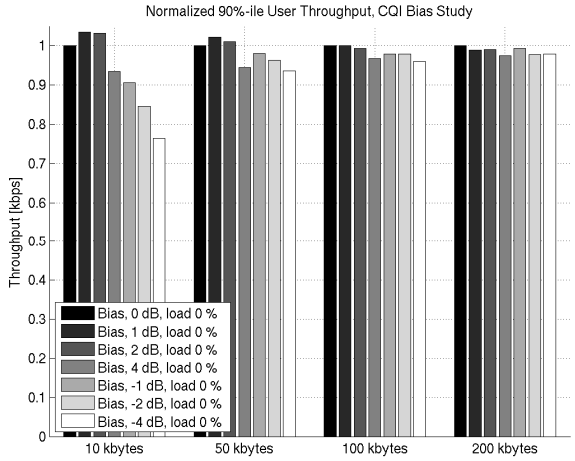


Figure 5. User Throughput, 90 Percentile, Fixed Bias

The reason why call length and amount packets affects to the fact that bias becomes more insignificant is that with longer calls OLLA has enough time to correct it. Thus, the impact of bias is directly proportional to the amount of OLLA iterations available during the call. With large file sizes the bias is corrected already when only e.g. 10% of file has been sent and then the rest of the file can be sent without the impact of bias. This naturally decreases the effect of the bias when e.g. throughput of the whole call is evaluated. This, behavior is confirmed by Figure 6. and Figure 7. which illustrate how well OLLA is able to converge the offset (CQI bias) when the call length is longer. Moreover, it should be also noted that even with ideal offset and low amount packets OLLA will not have time to converge in general as confirmed by distribution with higher amounts of source data. The relatively big amount of very low OLLA offsets in the end of the calls are results of UEs being in very good position where even the most aggressive MCSs are not enough aggressive.

B. Performance with random bias

The spectral efficiency with more realistic bias, which models the penetration levels of, e.g., different manufacturers' or differently configured terminals, is shown in Figure 8. In simulation environment realistic bias means bias which is randomized separately to each UE. As the figure shows, with random bias the performance is, expectedly, much more robust against the bias even with high range of bias values and low amount of data. Similarly to fixed bias study, with reasonable amount of source data the OLLA has enough samples and has time to fix the bias and thus it is not visible in SE. Moreover, it can be seen that the system load 0-100 % does not have noticeable impact to how bias impacts the performance.

Finally, first transmissions BLERs are illustrated in Figure 9. and Figure 10. As those figures show, if there is adequate amount of data the BLER operation point, which was assumed to be 20 %, in these simulations is maintained quite well, regardless of the bias. With lower amount of source data the OLLA operation point remains on higher level than the desired 20% target even without bias.

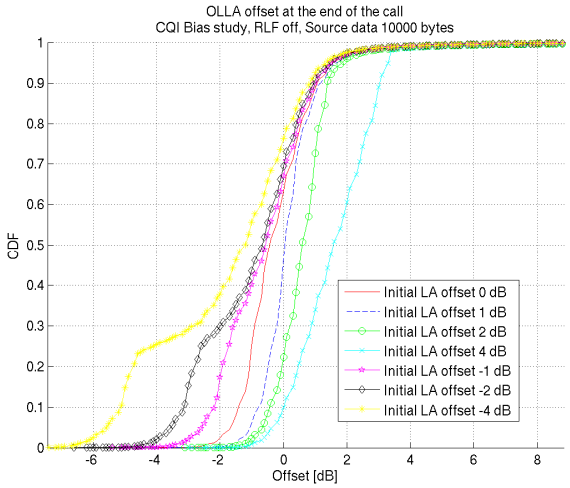


Figure 6. OLLA offset distribution, fixed bias, 10 kbytes source data

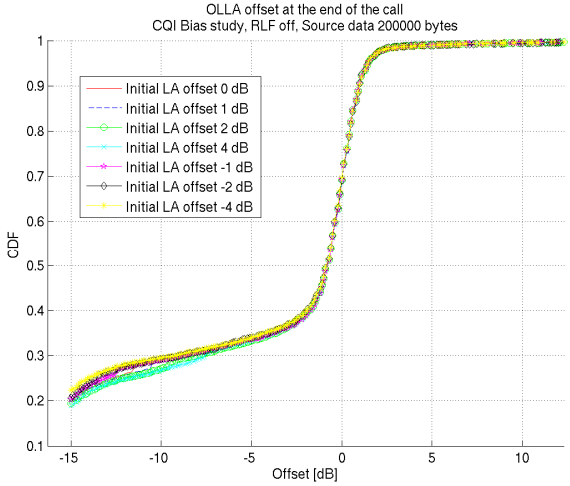


Figure 7. OLLA offset distribution, fixed bias, 200 kbytes source data

 ISBN:978-1-61208-113-7

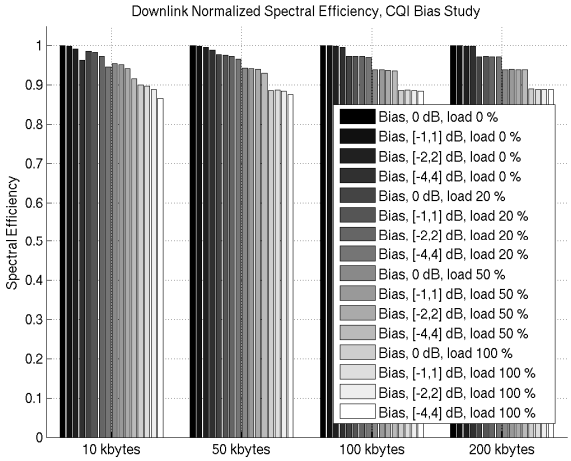


Figure 8. Normalized Spectral Efficiency, Random Bias

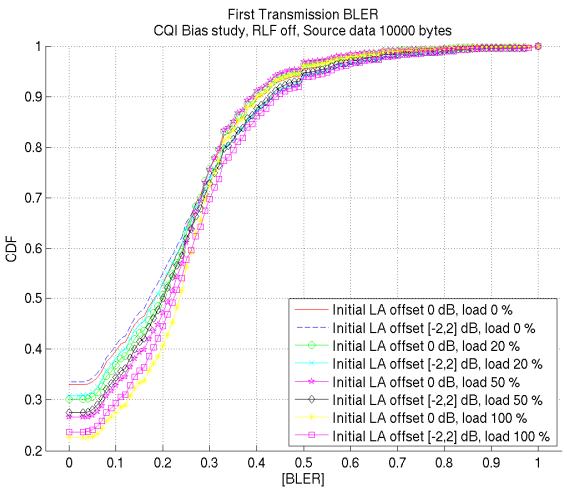


Figure 9. First transmission BLER per call, 10 kbytes data, Random Bias [-2,2] dB

VI. CONCLUSION

In this study we have shown with fully dynamic simulations how non-consistent CQI reporting by the UE impacts to the system level performance. Performance was evaluated with different combinations of traffic types, bias settings and bias values.

The results show that the system level performance can be affected by biased CQI values only when there is very low amount of data / packets and the bias is relatively high. In practice it is not, however, very likely that all of the users would have such a small amount of data and large bias. It is shown that OLLA is able to correct +-2 dB (random) bias range without affecting the spectral efficiency or user throughput but even higher bias values can be compensated with reasonable amount of data i.e. with high enough number of OLLA iterations.

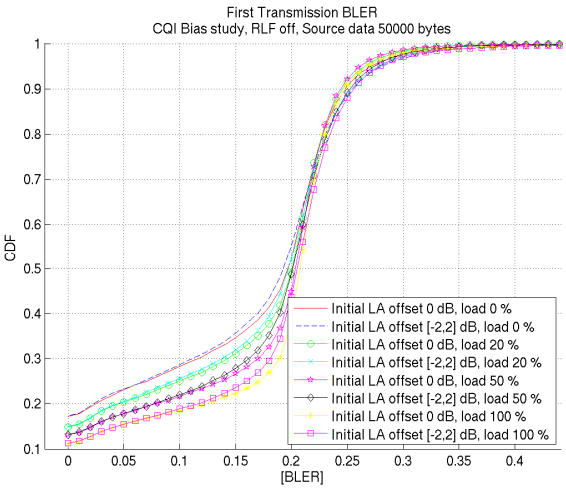


Figure 10. First transmission BLER per call, 50 kbytes data, Random Bias [-2,2] dB

ACKNOWLEDGMENT

This study is a collaborative work with Nokia and Nokia Siemens Networks. The authors would like to thank all of their co-workers and colleagues for their comments and support.

REFERENCES

[1] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, 3G evolution – HSPA and LTE for mobile broadband, 1st ed., Elsevier Ltd., 2007.

[2] H. Holma, and A. Toskala, LTE for UMTS – OFDMA and SC-FDMA based radio access, 1st ed., John Wiley and Sons Ltd., 2009.

[3] Technical Requirement Group, Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN), 3GPP TR 25.813, Rev. 7.3.0, March 2006.

[4] K.I. Pedersen, G. Monghal, I.Z. Kovacs, T.E. Kolding, A. Pokhariyal, F. Frederiksen, P. Mogensen, "Frequency Domain Scheduling for OFDMA with Limited and Noisy Channel Feedback", in Proceedings of IEEE Vehicular Technology Conference, 2007, VTC Fall 2007., September 2007.

[5] N. Kolehmainen, J. Puttonen, P. Kela, T. Ristaniemi, T. Henttonen, M. Moisio, "Channel Quality Indication Reporting Schemes for UTRAN Long Term Evolution Downlink", in Proceedings of IEEE Vehicular Technology Conference, 2008, VTC Spring 2008., May 2008.

[6] Technical Specification Group, User equipment (UE) radio transmission and reception, 3GPP TS 36.101, Rev. 8.6.0, June 2009.

 ISBN:978-1-61208-113-7

HyRA: A Software-defined Radio Architecture for Wireless Embedded Systems

Tiago Rogério Mück and Antônio Augusto Fröhlich
Federal University of Santa Catarina (UFSC)
Software/Hardware Integration Lab (LISHA)
Florianópolis - Brazil
{tiago, guto}@lisha.ufsc.br

***Abstract*—Traditional Software-defined Radio (SDR) architectures cannot go with the requirements of embedded systems, specially in terms of performance and power consumption. Low-power FPGAs now reaching the market might soon become a viable alternative to overcome such limitations. The *Hybrid Radio Architecture* (HyRA) introduced in this paper contributes to this scenario as it explores the *Hybrid HW/SW Component* concept to enable the implementation of SDRs as direct mappings of high-level synchronous data flow models. Although addressing SDR from a higher level of abstraction, HyRA mechanisms proved far more efficient than those behind GNU Radio when the target is an embedded reconfigurable hardware platform.**

Keywords- software-defined radio; embedded systems; FPGA.

I. Introduction

Wireless communication devices are at the heart of a growing number of embedded systems. Some of them, such as smartphones, must implement multiple communication protocols in face of constantly evolving standards. Others, such as wireless sensor network gateways, must simultaneously communicate under multiple protocols and sometimes even dynamically adapt themselves to preserve connectivity. In this scenario, *Software-defined Radio* (SDR) becomes an appealing approach, since most of the key components in the communication system—including the physical layer—are pushed into software, thus making them easy to reconfigure [1].

However, the implementation of a wireless communication system based only on an RF front-end, A/D converters, and a *General-Purpose Processors* (GPP) comes at a high cost. The associated *Digital Signal Processing* (DSP) algorithms demand very high processing power, a requirement that contradicts major design premises in the field, which usually include low cost, low energy consumption, and small size. Nevertheless, it is important to notice that this exceeding demand for processing power arises basically from the serialization of essentially parallel algorithms that takes place as they are pushed from hardware to software.

Implementing the key concepts behind an SDR on a reconfigurable hardware platform such as an *Field Programmable Gate Array* (FPGA) would preserve its main advantage—flexibility—without requiring a high-performance processor. For instance, an architecture based on DSP blocks on a datapath implementing a *Synchronous Data Flow* (SDF) could take advantage of the platform's inherent parallelism for the implementation of each individual DSP block and also to interconnect them efficiently. This has not been an option to embedded systems designers until now for a single reason: power consumption. Recent advances in low-power reconfigurable hardware, however, suggest that such systems can soon become viable. Indeed, several groups currently explore the use of hardware accelerators for the implementation of SDR algorithms [2]–[4]. *Single instruction, multiple data* (SIMD) extensions of GPPs, DSP processors, and functional blocks implemented in FPGAs are common approaches to limit processing power requirements at software level. Notwithstanding, the imminence of embedded SDRs fully implemented in low-power FPGAs calls for a systematic approach to guide the development of DSP components, interconnections, and controllers.

In this paper, we introduce HyRA, the *Hybrid Radio Architecture*, as a fundamental step toward a more comprehensive strategy to deploy SDRs in the context of embedded systems. HyRA relies on the *Hybrid HW/SW Component* concept of *Application-driven Embedded System Design* (ADESD) [5] to enable the implementation of an SDR as a direct mapping of a high-level SDF model. Each functional block in the model is associated to an hybrid component that can be plugged into HyRA's embedded SDR framework. Since hybrid components preserve their interfaces independently of how they are implemented, developers can freely decide which elements of the SDF graph go to software and which go to hardware. HyRA's framework features a programmable interconnect infrastructure that abstracts the *First In, First Out* (FIFO) channels between components. It also features a controller that dynamically coordinates the flow of data between components.

The remainder of this paper is organized as follows: Section II discusses related SDR implementation approaches; Section III recalls ADESD hybrid components, a fundamental concept behind HyRA; Section IV describes HyRA in details, while Section V presents an experimental evaluation; Section VI closes the paper with our conclusions.

II. Related work

The most straightforward SDR implementation approaches are the ones based on GPPs. These approaches target flexibility and ease of development, and usually delegate all processing to a GPP on a PC-like machine. These approaches

 ISBN:978-1-61208-113-7

are not suitable to embedded systems not only because of cost, energy consumption, and size, but also because of the overhead imposed by general-purpose operating systems and by the high-latency, high-jitter communication interfaces used to reach the RF front-end [6]. The GNU Radio [7] is the most representative case in this group. It features a framework and a library of signal processing blocks that enables SDRs to be built on ordinary PCs. In GNU Radio, the physical layer of a radio is abstracted as a flow graph in which nodes represent processing blocks and edges represent the data flow between them.

Another approach is to delegate signal processing to programmable devices specifically designed for that purpose. Several architectures [2]–[4] relies on a GPP processor coordinating multiple DSP processors with SIMD and *very long instruction word* (VLIW) datapaths. Some architectures, such as the Elemental Computing Architecture [8], define fine-grained components specific to a given class of operations which can be configured and connected to each other to build an SDR. Differently from HYRA, these approaches focus on the efficient implementation of individual DSP blocks, without addressing the relationship between the implementation and a high level model. Also, the resource allocation and synchronization of processing elements must controlled manually by the programmer. Some tools and languages, such as SPIR [9], aim to provide means to compile high level models of DSP applications into code that is suitable to run on *multiprocessor System-on-Chip* (MPSoC) DSP architectures. This is an important step toward a higher level SDR development strategy, but, as authors recognize, some algorithms that require considerably more processing power than the average, such as filters, searchers, and Turbo decoder, easily become bottlenecks in the programmable DSP hardware approach.

Apart from the software-based approaches, the dedicated implementation of the wireless communication protocols in FPGAs is another common approach to put together the flexibility of a reconfigurable radio, and the efficiency of a dedicated hardware. However, implementing a complex digital hardware design is not a straightforward task. Even if tools to translate high level specifications to a synthesizable *register transfer level* (RTL) description exist [10], [11], there is still a lack of means to integrate the dedicated hardware dataflow in a control flow that also encompasses software processes on the GPP. As a result, any change in the SDR protocol that requires more than a change on the parameters of existing hardware blocks usually requires the generation of a new hardware instance.

III. ADESD HYBRID COMPONENTS

HYRAwas developed based on the idea of *hybrid hardware/software components* [5]. This concept is an elaboration on the concept of hardware mediators proposed in the *Application-driven Embedded System Design* (ADESD) [12] methodology. In ADESD, hardware mediators are a particular kind of component that are responsible for keeping the high-level abstractions independent of the hardware platform. This components are implemented using *generative programming* techniques, adapting the hardware interface to the interface required by the system instead of creating a hardware abstraction layer. The idea of *hybrid hardware/software components* emerges from the fact that different mediators can exists for the same hardware component, each one designed for different purposes (e.g., changing the trade-off between performance and energy consumption). Each component aggregates mediators for its many implementations, which could be in hardware, software or both. According to system requirements of cost, performance, energy, etc., any one of these implementations can be selected without any change to the higher system layers that use the component.

In previous works [5] were defined and implemented in the *Embedded Parallel Operating System* (EPOS) [12] some architectural guidelines for the translation of operating system related components, such as timers, schedulers, and synchronizers, from software to hardware and vice versa. Whether such guidelines can also be defined for DSP related components has not yet been investigated, but nonetheless, a hybrid component is a convenient construct to encapsulate functional blocks on an SDR. This will be demonstrated in the next sections.

IV. SDR IMPLEMENTATION WITH HYRA

HYRA relies on SDF abstractions of SDRs. In this model, the SDR processing chain is abstracted as a flow graph, where the nodes represent processing blocks and the edges represent the data flow between the blocks. In HYRA, each functional block in the SDF is associated to a hybrid component that can be plugged into HYRA's embedded SDR framework. The developer uses this framework to specify connections that defines the data flow between components. The framework is responsible for creating the FIFO channels between the components and for starting the runtime mechanism that dynamically coordinates the data flow.

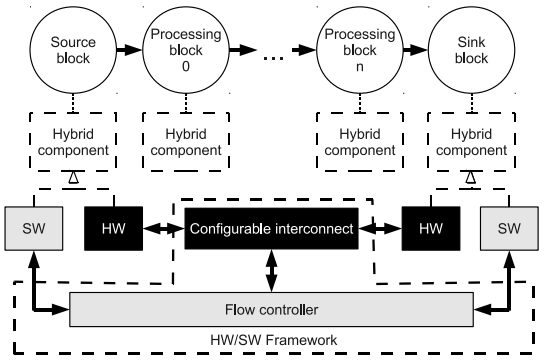


Fig. 1. Overview of HYRA

Figure 1 shows an overview of our architecture. It's framework have both a hardware and software side. The hardware side features a configurable interconnect structure which provides a FIFO-like stream interface to interconnect the hardware implementations of hybrid components. Also, it offers the

resources necessary to create HW FIFO channels between components in hardware and to coordinate their execution. The software side offers the interfaces to connect the software implementations of hybrid components and the runtime mechanism, denominated *Flow Controller*, which is responsible for controlling the connections and synchronization between components. The following sections will explain in more details each part of the architecture.

A. Flow controller

The *flow controller mechanism* is responsible for connecting the components and controlling the data flow between them at runtime. When the developer specifies a connection between two components, the flow controller creates a FIFO channel between them. The size of the FIFO in the channel is defined by the following equation:

$$FIFO_{size} = max(Blk_0^{outputrate}, Blk_1^{inputrate}) \cdot \alpha \quad (1)$$

in which an output of Blk_0 is being connected to an input of Blk_1, $Blk_0^{outputrate}$ is the number of data elements generated upon each execution of Blk_0, and $Blk_1^{inputrate}$ is the number of data elements consumed in each execution of Blk_1. The FIFO size is formulated in this way based on the fact that Blk_0 cannot generate data faster then Blk_1 can consume. If this happens in the system, due to modeling error or poor performance of Blk_1, the FIFO will always overflow. α is a safety factor that should be set according to the jitter characteristics of the platform.

The FIFO allocation will depends on the actual physical implementation of the hybrid components that the channel is connecting. If both are implemented on software, a SW FIFO will be dynamically allocated in the system main memory. If one or both components are in hardware, the *flow controller mechanism* will allocate the FIFO inside the hardware interconnect structure. The next section will explain the hardware side of the framework.

The control of the data flow between software components is accomplished by creating a thread for each component. Each thread executes a loop where, at first, it remains locked onto semaphores associated with the channels connected to the block's inputs. Each time an element is added to a channel, the *v()* method of its associated semaphore is called, unlocking the threads that consume the data from the channels. After acquiring all the semaphores, the thread consumes the inputs, executes the block's processing, and finally writes the result in the output channels, unlocking the threads associated to the subsequent blocks.

B. Hardware support

Hybrid components implemented in hardware don't use the software synchronization mechanism described in the previous section. Instead, they are controlled directly by signals provided by the FIFO channels in hardware. The deployment of HW FIFO channels is supported by the flow controller hardware structure shown in Figure 2. This structure mainly consists of a interconnection block that have a set of read ports, write ports and internal FIFOs, where the connection between these three elements can be defined by software-controlled configuration registers.

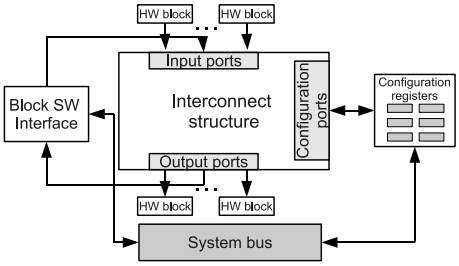


Fig. 2. HW layer of HYRA's framework

All the components that are implemented as hardware have their inputs connected to the structure's read ports, and their outputs connected to the write ports. When these components are connected, the flow controller mechanism uses the information provided by the component's software interface to define which port must be connected to which FIFO. Since the HW FIFOs must have a fixed size, the interconnect structure allows FIFOs to be interconnected among them. This way, when two components are connected, it is possible to allocate a chain of FIFOs between them, in a way in which the total size of the chain is bigger than or equal to the required FIFO size.

Figure 3 shows how we have implemented this interconnect structure. We used a simplified butterfly fat tree NoC architecture [13] optimized for the interconnection of stream blocks. It consists of a matrix of FIFOs where each FIFO input is connected to each input port, and each output port is connected to each FIFO output. Each FIFO output is connected to the input of the FIFO in the next column on the same line. With this interconnect scheme we can provide a wide range of possible allocation for each input/output port, while keeping the use of FPGA resources by interconnect at a reasonable level.

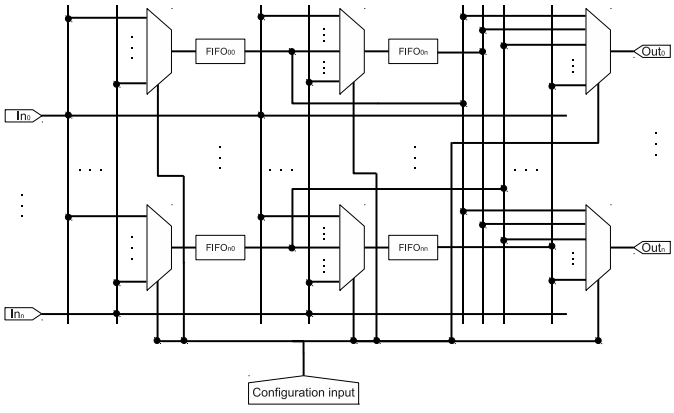


Fig. 3. Overview of the flow controller HW interconnect internal structure

In order to provide all possible connections between HW

 ISBN:978-1-61208-113-7

and SW components, four different FIFO channel implementation are required: *SW FIFO*, *HW FIFO*, *SW-HW FIFO*, and *HW-SW FIFO*. We already explained how channels between SW-only and HW-only components are created. The connection between software and hardware components is achieved through the *Block SW Interface* shown in Figure 2. It behaves like a wrapper between the system bus and the interconnect structure stream interface. When a component in hardware is connected to a component in software, its respective ports are connected to ports associated with the *Block SW Interface*. When there is a SW→HW connection a *SW-HW FIFO* provides a software interface so the source block can write to the *Block SW Interface* output port associated to the destination block input port. But, the HW→SW connection requires additional runtime and hardware support. When a *HW-SW FIFO* is created, it register itself in the interrupt handler for the *Block SW Interface*'s interrupts. Every time new data arrives at one of the FIFOs connected to the *Block SW Interface*'s input ports, it will issue an interrupt that will release the semaphore associated with the FIFO, as described in the previous section.

V. Evaluation and results

In this work, we have focused only on the architectural support and data flow aspects for the implementation of SDR. In order to evaluate the proposed architecture we disconsidered the signal processing function, since they are coved in other works [10], [11], and focus only on HyRA's intrinsic overhead. We have evaluated this overhead in two aspects: area overhead (FPGA resource utilization) and performance overhead (latency added to the data flow by the hardware and software control structures).

A. Evaluation setup

To define the data flow structures for our evaluations, we have analyzed the data flow structure of the physical layer of several protocols covering a wide range of modulation schemes and application classes: Bluetooth, UWB, ZigBee, Wi-Fi (802.11a), and W-CDMA. In this analysis we verified that all protocols follows a common data flow structure. On the receive chain, there is usually a filter before the demodulation blocks, normally a low pass filter used to obtain a clean piece of spectrum that contains the information. Next, there are the demodulation/synchronization blocks, which normally consists of one or more data flows being processed in parallel. The last step is a post-demodulation filter which normally consists of a channel decoder for error detection and correction. The transmit chain follows an analogous structure. From this general structure, we have defined the structures shown in Figure 4 to evaluate the overhead in terms of the number of blocks in a data flow (4a) and the number of data flows in parallel (4b), covering many possible variations of the general structure described previously. We also analyzed how the number of inputs/outputs of a block affects the overhead (4c).

These structures are composed by three kinds of blocks. The *Timestamp source* block generates samples which consist of timestamps that represent the time when the sample was generated. The *dummy blocks* are empty blocks that just propagate their inputs to their outputs. After being generated, the samples will go through the *dummy block* chain. When a sample arrives at the *Timestamp sink* block, the timestamp is compared with the current time, obtaining the time the sample took to go through the *dummy block* chain. Since the *dummy blocks* are empty, this resulting time represents only the overhead imposed by the architecture on the data flow. There is also the *Fork block* and *Join block* which are used to fork and join the data flows, respectively.

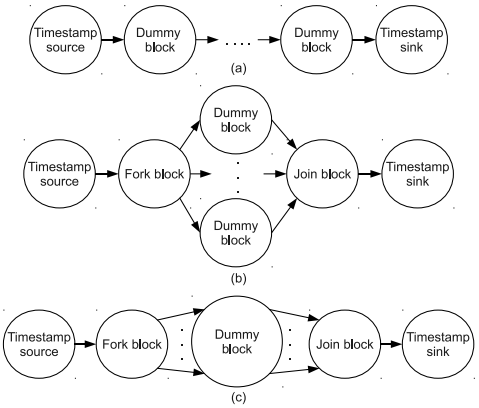


Fig. 4. SDRs data flow structures used for the overhead evaluation in terms of the number of blocks in serial (a), the number of data flows in parallel (b), and the number of inputs/outputs (c)

B. System implementation and configuration

To evaluate these three structures, we have implemented HyRAon the EPOS operating system running on the Xilinx's ML403 Embedded Platform. The ML403 features a Virtex-4 FPGA with an embedded PowerPC 405 microprocessor. In order to use the same hardware configuration, we have synthesized the hardware with all of the necessary blocks for all experiments. We have used the following tools and parameters: ISE/EDK 10.1; GCC 4.0.2; FPGA and microprocessor clocked at 100 MHz; interconnect structure configured with 32 input/output ports, and 64 FIFOs (8 bit wide with 16 elements); the α factor fixed to 1 in order to provide an evaluation considering low jitter requirements.

Table I shows the resource consumption of the generated hardware. Separate results are shown for HyRA's structures along with the HW dummy blocks, and for the system IPs generated by EDK (internal memory, memory controller, interruption controller, UART, etc). Our architecture alone uses about 65% of the available logic and 0% of the available memory. Due to the lack of memory blocks available on the device, we chose to implement the FIFOs using the SRL16 capabilities to convert a 4-input LUT into a 16-bit shift register. Howsoever, this apparently high resource usage is due to the very limited amount of logic available on the used device. When compared to other system IPs, we can see that HyRAuses slightly more resources then a complete set of basic IO and memory peripherals.

 ISBN:978-1-61208-113-7

TABLE I
AMOUNT OF HW RESOURCES USED BY THE SYNTHESIZED STRUCTURES

Resource	Our IPs	EDK IPs	Full System
4-input LUTs	37%	35%	72%
Slice Flip Flops	67%	31%	98%
Occupied Slices	70%	55%	99%
RAM blocks	0%	63%	63%
Max. frequency	167 MHz	109 MHz	107 MHz

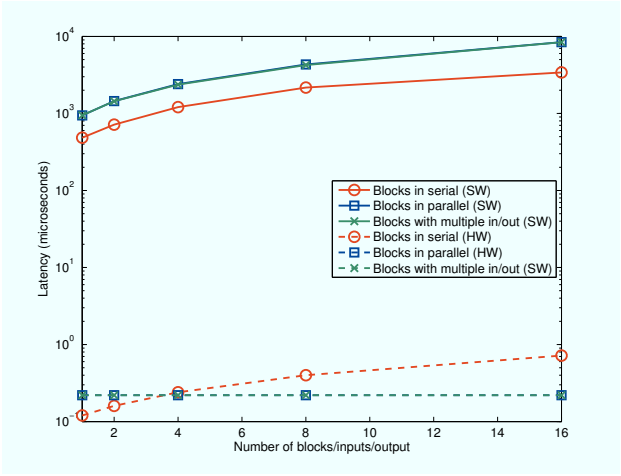
C. Performance overhead

We have done tests to determine the performance overhead that we have defined as the latency between the *Timestamp source and sink* blocks in the three basic data flow structures. We have implemented each structure in hardware and software and executed tests with the number of dummy blocks ranging from 1 to 16. In each test 6×10^7 samples were generated and we obtained the average value of the latencies of each sample and the standard deviation which was used to obtain the coefficient of variation. A sampling rate of 1×10^6 samples/second was used in the tests with blocks in hardware. For the tests with blocks in software we used a sampling rate of 1×10^4 due to the low speed of the PowerPC processor.

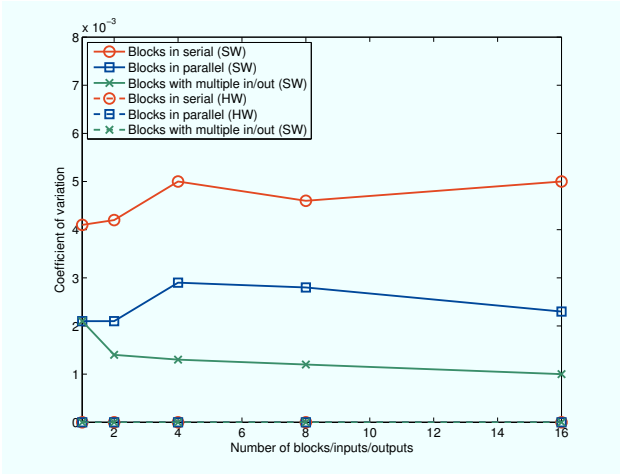
Figures 5a and 5b show the results. When using only software blocks, the overhead grows linearly in relation to the increase in the number of blocks and the number of inputs and outputs for all structures. The coefficient of variation remained low in all configurations. When using only HW blocks, the latency was about four orders of magnitude lower than when using software blocks and, as expected, except for the serial block configuration, the latency remained constant regardless the size of the structure, due to the full parallelism that can be explored in this kind of architecture. There is also a null coefficient of variation in the hardware operations.

To evaluate the communication latency between components implemented in HW and components implemented in SW, we have used the data flow shown in Figure 6 which cover operations on both *SW-HW FIFOs* and *HW-SW FIFOs*. We have performed the same experiment described previously on this two structures and verified that the average latency on both interleaved data flows yielded similar results: $221\mu s$ and $234\mu s$ for data flows *(a)* and *(b)*, respectively. Data flow (a) have more SW blocks then (b), thus showing a higher SW management overhead. However, both have two *SW-HW FIFOs* and two *HW-SW FIFOs* connecting the blocks. By comparing the latencies with the ones obtained for SW-only blocks ($221\mu s$) and HW-only blocks ($0.31\mu s$) we can see that the read/write operations SW channels represents the most significant overhead.

We also compared the overhead of our architecture to GNU Radio. For this comparison, we replicated the same tests described previously using GNU Radio running over a Linux operating system in a PC. Our architecture and EPOS were compiled for the IA32 architecture only with software blocks support, and evaluated in the same system. For the GNU Radio experiment, we used GNU Radio 3.2.2 running on a Linux kernel 2.6.28. The result for the serial blocks data flow structure shown in Figure 7a demonstrates that our architecture performance surpasses GNU Radio between 2 and 4 times, and this difference increases as the number of blocks in the processing chain increases. Figure 7b also shows that we are able to achieve smaller latency variations as well.



(a) Average latency



(b) Coefficient of variation

Fig. 5. Latency for blocks in serial, in parallel, and with multiple input/output

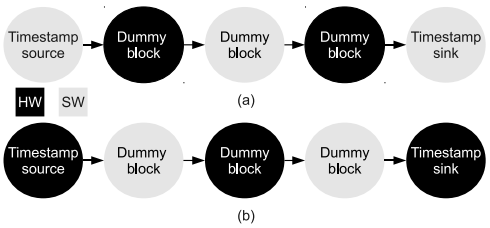
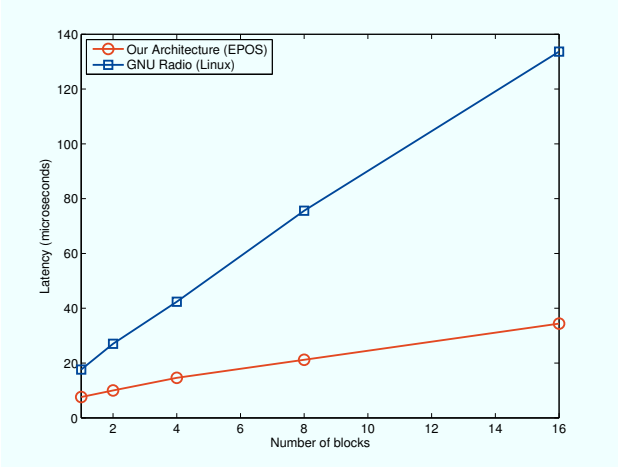


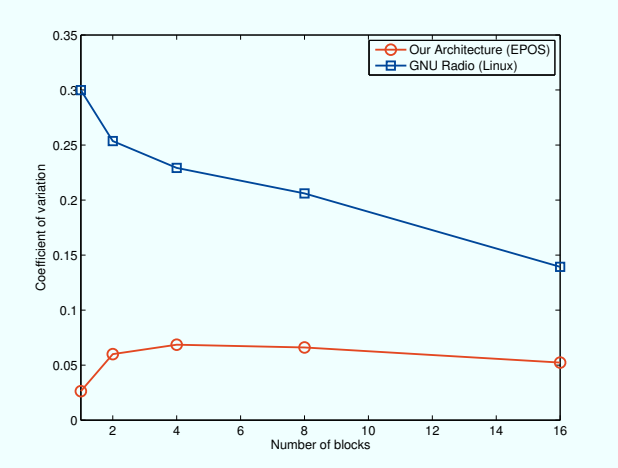
Fig. 6. Serial data flows with interleaved SW and HW blocks

D. Discussion

The results show that our architecture yields superior performance than an equivalent, in terms of abstraction level,



(a) Average latency



(b) Coefficient of variation

Fig. 7. Latency of the proposed architecture VS GNU Radio on the serial blocks data flow structure

commonly used architecture. Only software components were used in this comparison, if a hardware device was used to generate the timestamps, GNU Radio would suffer an additional disadvantage. In GNU Radio, the use of any hardware device to obtain or sink data from/to the environment requires Linux drivers whose performance is mostly limited by the kernel's abstraction layer. A previous work [6] shows that, due to the Linux kernel overhead, the standard deviation of the time a sample takes to get to the processing chain after being generated in the RF Front-end is higher than the average time. This problem does not appear in EPOS since the metaprogrammed hardware mediators are dissolved within the application when the system is compiled, which leads to higher performance.

However, even with known latency problem, the GNU Radio is widely used and several protocols have been successfully implemented using it. The results have shown that with our architecture we were able to bring similar functionality with superior performance to the embedded system domain, which leads to the conclusion that our architecture is suitable for the implementation of high-end protocols in embedded systems.

VI. Conclusion

In this paper we have introduced HyRA, an *Hybrid Radio Architecture* that explores the *Hybrid Component* concept within ADESD to enable the implementation of SDRs as direct mappings of high-level SDF models. As hybrid components, HyRA SDR blocks can be implemented as arbitrary combination of software and hardware on FPGA-based platforms. The programmable interconnect infrastructure in HyRA's framework ensures transparency in this respect. FIFO channels can be fine tuned to fulfill the requirements of a given SDR protocol, while the controller dynamically coordinates the flow of data between components.

In comparison with other approaches, HyRA addresses the implementation of SDRs in the context of embedded systems from a higher level of abstraction. Moreover, the evaluation results presented in this paper confirm that the overhead caused by the proposed architecture in terms of latency is much smaller than that of GNU Radio, a widely accepted architecture. Furthermore, our experiments demonstrated that HyRA can be implemented on reconfigurable hardware platform with minimal additional resources. In combination, this factors confirm that our architecture meet the requirements for the implementation of high-end protocols in embedded systems.

References

[1] E. Buracchini, "The software radio concept," *IEEE Comm. Mag.*, vol. 38, no. 9, pp. 138–143, 2000.

[2] J. Glossner, E. Hokenek, and M. Moudgill, "The Sandbridge Sandblaster Communications Processor," in *3rd WASP*, 2004, pp. 53–58.

[3] M. Woh, Y. Lin, S. Seo, S. Mahlke, T. Mudge, C. Chakrabarti, R. Bruce, D. Kershaw, A. Reid, M. Wilder, and K. Flautner, "From SODA to scotch: The evolution of a wireless baseband processor," in *MICRO 41*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 152–163.

[4] K. van Berkel, F. Heinle, P. P. E. Meuwissen, K. Moerman, and M. Weiss, "Vector processing as an enabler for software-defined radio in handheld devices," *EURASIP*, vol. 2005, pp. 2613–2625, 2005.

[5] H. Marcondes and A. A. Fröhlich, "A Hybrid Hardware and Software Component Architecture for Embedded System Design," in *IESS '09*, Langenargen, Germany, 2009, pp. 259–270.

[6] G. Nychis, T. Hottelier, Z. Yang, S. Seshan, and P. Steenkiste, "Enabling MAC Protocols Implementation on Software-defined Radios," in *NSDI '09*, 2009.

[7] GNU FSF project, "The GNU Radio," 2010, [retrieved, November 8, 2010]. [Online]. Available: http://www.gnu.org/software/gnuradio

[8] Steven Kelem, Brian Box, Stephen Wasson, Robert Plunkett, Joseph Hassoun, and Chris Phillips, "An Elemental Computing Architecture for SD Radio," in *SDR '07*, 2007.

[9] Y. Lin, M. Kudlur, S. Mahlke, and T. Mudge, "Hierarchical coarse-grained stream compilation for software defined radio," in *CASES '07*. New York, NY, USA: ACM, 2007, pp. 115–124.

[10] Z. Guo, W. Najjar, and B. Buyukkurt, "Efficient hardware code generation for FPGAs," *ACM TACO*, vol. 5, no. 1, pp. 1–26, 2008.

[11] F. Plavec, Z. Vranesic, and S. Brown, "Towards compilation of streaming programs into FPGA hardware," in *FDL '08*, 2008, pp. 67–72.

[12] A. A. Fröhlich, *Application-Oriented Operating Systems*, ser. GMD Research Series. Sankt Augustin: GMD - Forschungszentrum Informationstechnik, Aug. 2001, no. 17.

[13] P. P. Pande, C. Grecu, A. Ivanov, and R. Saleh, "Design of a switch for network on chip applications," in *ISCAS '03*, vol. 5, 203, pp. V217–V220.

 ISBN:978-1-61208-113-7

Sensors Deployment Strategies for Rescue Applications in Wireless Sensor Networks

Inès El Korbi, Nesrine Ben meriem, Leila Azouz Saidane
CRISTAL Laboratory
National School of Computer Science
University of Manouba, Tunisia
e-mails: ines.korbi@ensi.rnu.tn, bmariem.nesrine@gmail.com, Leila.saidane@ensi.rnu.tn

Abstract— **Wireless sensor networks (WSNs) become a major tool for various security and surveillance applications to detect and monitor environmental changes, to control vehicle traffic, etc. For all these capabilities, we propose in this paper to use WSNs in rescue applications. These applications are critical in terms of network response time when disaster strikes. The network response time is the time required by sensor nodes to detect victims' positions and forward them to the sink node. In this paper, we consider two sensor nodes deployment strategies: linear and circular strategies. By deployment strategies, we mean initial nodes positioning and nodes movement procedures. Therefore, we simulate these two strategies under the WSNet simulator and derived the Monitored Area Sweep Time (MAST) (i.e., the time required by sensor nodes to scan all the area coordinates to find victims) and the energy dissipation due to node mobility. Simulation results are verified using analytical expressions of the zone sweep time and the energy dissipation. Finally, we extend the behavior of the sensor nodes deployment strategies to support nodes communication so that mobile sensors can forward victims' positions to the sink node.**

Keywords- wireless sensor network; mobility; sensor nodes deployment strategies; rescue applications.

I. INTRODUCTION

Recent advances in miniaturization such a low power circuit design and low powered wireless communications make it possible the emergence of a new kind of equipments: sensor nodes. The sensor node is a few cubic centimeters device with various capabilities such as simple wireless communication, minimal computation facilities, and sensing of the physical environment. Typical sensing tasks are temperature, light, vibration, sound, radiation, etc. The above characteristics of sensor nodes, allowed the usage of Wireless Sensor Networks (WSNs) [3], [13] in different fields on applications.

Indeed, WSNs were initially designed for military applications, such as battlefield surveillance and enemy tracking. Now, they are used in many industrial and civilian application areas, such as industrial process monitoring, environment and habitat monitoring, healthcare applications and traffic control [10].

In this paper, we focus on area monitoring applications and particularly on rescue applications to detect victims' positions when hazardous events occur (seism, fire, explosion, etc.) [5], [11]. This kind of applications imposes strict time requirements and the efficiency of WSN deployment procedure depends on the network response time. The network response time corresponds to the time required by the sensors deployed in the monitored area to detect victims' positions and notify the sink node accordingly. Therefore, the faster, the victims are localized, the faster rescue personal can perform their tasks and help victims. In our work, we focus on sensor nodes with motion capabilities. Mobile sensors are initially deployed over the monitored area. They remain in their fixed positions until receiving a hazardous event message from the sink node that triggers their movement and the victims' detection procedure.

Therefore, we propose in this paper, two sensor nodes deployment strategies to detect victims in the wireless network. These techniques are called circular and linear strategies. For these two techniques, we propose algorithms to both initially place sensor nodes and then moving them within the monitored area. Then, we simulate both linear and circular strategies using the WSNet simulator [14]. From simulation results, we derive the Monitored Area Sweep Time (MAST) which corresponds the time required by the sensor nodes to scan all the monitored area coordinates to localize victims. Then, we evaluate the energy consumption due to node mobility for both techniques. Simulation results are verified analytically using both sweep time and mobility based energy consumption expressions. Analytical results verify that the nodes initial deployment and movement algorithms are correctly implemented under the WSNet simulator. Hence, these algorithms could be extended to support communication aspects between sensor nodes for victim detection purposes.

Indeed, in the last part of the paper, we focus on victim detection mechanism by introducing a message exchange procedure between the different sensors within the network to localize victims and communicate their positions to the sink node. We then evaluate the energy consumption due to both communication and mobility and compare the MAST value to the last victim detection time (the time elapsed until the last victim in the monitored area is detected) for different sizes of the monitored area.

In the rest of the paper, we review in Section 2 the works related to our study. Section 3 presents the concepts and the algorithms of the sensors nodes deployment strategies. In Section 4, we simulate our strategies using the WSNet network simulator [14] and derive expressions of the monitored zone

 ISBN:978-1-61208-113-7

sweep time. We also derive energy consumption curves due to sensors mobility. Simulation results are verified analytically using expressions of both sweep time and energy consumption due to mobility. Section 5, presents the victim detection mechanism within the monitored area by ensuring communication between the sink node and the mobile sensors. We therefore derive the resulting energy consumption due to both mobility and communication and compare the MAST value to the last victim detection time. We conclude the paper and present our future work in Section 6.

II. RELATED WORKS

Area monitoring applications use WSNs either for measuring and surveying purposes or for reporting various types of activities and events. Therefore, the area coverage criterion has to be met. A point is covered by a sensor if it is within its sensing range. In [1], authors design a distributed self deployment algorithm for coverage calculations in mobile sensor networks and consider various performance metrics, like coverage and uniformity. The work in [15] assumes that a cluster head is available to collect information and determine the target location of the mobile sensors. Sensor deployment has also been addressed in the field of robotics [8], [9], where sensors are deployed one by one, utilizing the location information of previously deployed sensors.

After being initially deployed over the monitored area, sensor nodes have to sweep the monitored zone to detect victims. Sweeping algorithms have received much attention in the past few years. The performance of these algorithms can be evaluated from various aspects, including the achieved coverage percent, the number of deployed sensors and the time required for the sweeping. In [2], authors investigate the problem of how to optimally move mobile sensors lying within a region to the perimeter of that region to detect intruders. In [6], Rekleitis et al. propose an approach to sweep all the destination zones. They assume that mobile nodes can communicate with each other and know their positions. The area is divided into stripes, with each mobile node taking care of one stripe. Wong et al. [12] use topological mapping to sweep the destination area. They make cell decomposition and cover each cell by a zigzag pattern. Batalin and Sukhatme [7] propose a decentralized method and present the frequency coverage metric to evaluate the quality of sweep coverage. The common challenges of these studies are the coverage ratio and energy consumption. Therefore, this paper aims to develop two efficient sweeping algorithms (the linear and circular sweeping algorithms) such that the number of deployed sensor is as few as possible while the monitored region is guaranteed to have full coverage.

In the next Section, we introduce two sensor nodes deployment techniques where sensors can move autonomously to sweep the monitored area.

III. SENSOR NODES DELPLOYMENT STRATEGIES

In this Section, we define two sensors deployment strategies to localize victims in rescue applications. The deployment strategies define at the same time the initial nodes locations and nodes movement when a hazardous event occurs. These two components of the sensors deployment strategies will definitely determine the efficiency of a strategy and its response time to perform victim localization. We first define concepts of linear and circular strategies.

A. Fundamental concepts of linear and circular strategies

When deploying sensor nodes within the monitored area, two basic criteria have to be met:

- At any time, communication between every sensor node and the sink node has to be maintained.
- All the points of the monitored area have to be covered by the wireless sensors.

Therefore, we define the following quantities:

- R_S : The node sensing range
- R_C : The node communication range, $R_C = 2R_S$
- N : The number of mobile sensors in the monitored area.
- p : The sink node placed at the center of the monitored zone and communicates victims' locations to the centralized system. The sink node's coordinates are initially know and given by (x_p, y_p). Fixing the sink position will simplify the communication mechanisms between the sink and the mobile sensor nodes.
- D : The diagonal of the zone to be monitored by the sensors. In the rest of the study, we consider that the monitored zone is square shaped. This assumption can be easily extended to a rectangular zone with height x and width y , where $D = y\sqrt{2}$ (and $y = \max(x, y)$).

1) The linear strategy

The linear strategy illustrated in Figure 1 consists in placing the wireless sensor nodes on axes horizontally and vertically.

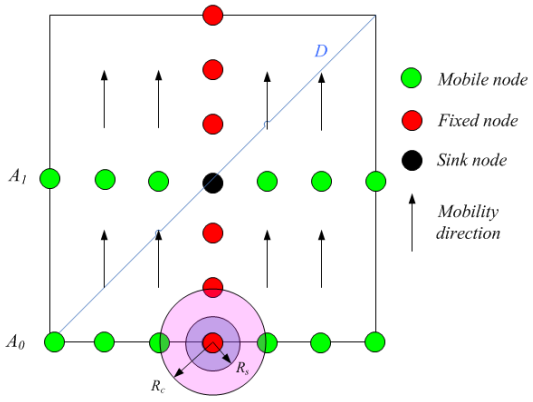


Figure 1. Sensor nodes initial positions according to the linear strategy

Therefore, we define two sensor categories:

- *Fixed sensors*: are placed vertically on the monitored area aligned with the sink node on its both sides to

keep communication with the mobile nodes when they are moving.

- *Mobile sensors*: are placed on horizontal axes such as the distance that separates two adjacent sensors on a horizontal axis is always R_c. When an axis is saturated, we place another horizontal axis parallel to the previous axis.

The number of horizontal axes depends on the number of sensors to be placed on the monitored area.

Horizontal axes are numbered from A_0 to A_{K-1}, where K is the number of mobile axes. Axis A_0 sensors are situated on the lowest boundary side of the monitored area (Figure 1). When an event happens, the mobile nodes move along vertical axes in bottom up direction until sensors on a given horizontal axis A_i reach the position of their successors' nodes on the next axis A_{i+1}. We say that $n_{i+1,j}$ is $n_{i,j}$ successor's node if $n_{i,j}$ is placed on axis A_i and $n_{i+1,j}$ is placed on axis A_{i+1}, such as $n_{i,j}$ and $n_{i+1,j}$ have the same x_j abscissa. Nodes stop moving when the nodes on axis A_{K-1} reach the upper boundary side of the monitored area.

When moving along vertical axes, a node situated on a given horizontal axis will forward the information of a victim location to its neighbor on the same horizontal axis until its message reaches one of the fixed sensors that will forward the received message vertically till it reaches the sink node. As in [11], we assume that the sensors are location aware (i.e., they know their geometric coordinates when they are initially deployed). When moving, the new positions are determined as a function of the old ones.

2) The circular strategy

The circular strategy consists to initially place the N mobile sensors around the sink node on axes as in Figure 2. The distance separating two adjacent nodes on the same axis is equal to R_c. Each node on a given axis defines a level.

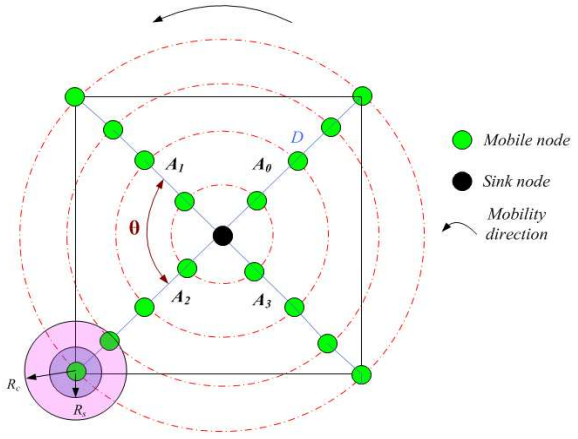


Figure 2. Sensor nodes initial positions according to the circular strategy

A sensor node is said to be on level i, if the distance that separates this node from the sink is equal to $R_c \times i$.

Two adjacent nodes on a level i are separated by a distance equal to $R_c \times \theta$, where θ is the angle between two adjacent axes. θ is equal to $2\pi / K$, where K is the number of axes within the network.

When a hazardous event occurs within the sensor network, each sensor node turns around the sink in a counter-clockwise way. Sensors on the same level i turn around the sink simultaneously at the same speed. Moreover, nodes on the same axis (belonging to different levels) have to progress together to keep communication with the sink node. Hence, when moving, nodes have to be arranged along their original axes at any point of time. To keep axes structure, two nodes $n_{i,l}$ and $n_{j,l}$ belonging to the same axis A_l and placed on levels i and j have to perform the same angular distance θ_r, between two time instants 0 and t_r where θ_r is given by:

$$\theta_r = \frac{D_i}{R_i} = \frac{D_j}{R_j} \quad (1)$$

where D_i (respectively D_j) is the distance (in meters) traversed by a sensor on level i (respectively on level j) and R_i (respectively R_j) is the distance separating the node $n_{i,l}$ (respectively $n_{j,l}$) from the sink. Hence, according to equation (1), D_j will be equal to:

$$D_j = \frac{R_j}{R_i} D_i \quad (2)$$

As the two nodes $n_{i,l}$ and $n_{j,l}$ have to go through distances D_i and D_j within the same time interval $[0, t_r]$, the velocity of node $n_{j,l}$ have to be $\frac{R_j}{R_i}$ times the one of node $n_{i,l}$. Finally, sensor nodes stop moving, when each sensor performed an angular distance equal to θ.

If a victim was detected while moving around the sink node, a mobile node $n_{i,l}$ located on axis A_l and level i, will forward the victim's coordinates to the node $n_{i-1,l}$ located on the same axis A_l and the lower level $i-1$. This procedure is repeated until the victim coordinates reach the sink node.

B. Proposed algorithms for linear and circular strategies

1) The linear strategy algorithms

In the following, we detail the algorithms corresponding to the initial nodes positioning and nodes movement algorithms using the linear strategy. Algorithms related to nodes communication to detect victims' positions and forward them to the sink node are detailed in Section 5. We define:

- N_f : The number of fixed sensors. It also corresponds to the number of sensors per mobile axis.

 ISBN:978-1-61208-113-7

- K: The number of horizontal axes.
- V : The velocity of a mobile node.

Algorithm 1: Initial Sensor Nodes Positioning

Input: N , D , V , R_c and (x_p, y_p)
Output: *NList*: list of mobile and fixed sensor nodes

```
1:  N_f ←ceil( D /(2*sqrt(2)*R_c ))*2;
2:  K←floor( (N – N_f )/ N_f )
3:  Sn←new Node(); Sn →x ← x_p ; Sn →y ← y_p
4:  Add(Sn, NList);
5:  h1← -1;  h2← -1
6:  //Deploying  fixed nodes
7:  For i from 1 to N_f do
8:        Sn ←new Node(); h2← h2*h1
9:        Sn →x ← x_p ; Sn →y ← h2* Rc*i
10:       Sn →speed ← V ; Add(Sn, NList);
11: End
12: //Deploying  mobile nodes
13: For i from 1 to K do
14:   For j from 1 to N_f do
15:       Sn ←new Node();
16:       h2← h2*h1
17:       Sn →x ← x_p + h2*j* R_c
18:       Sn →y ← D/( sqrt(2) *i)
19:       Sn→speed ← V ; Add(Sn, NList);
20:   End; End
```

Since we consider floor and ceil functions, only $N_f \times (K+1)$ sensor nodes will be deployed on fixed and mobile axes and participate in the victim detection procedure. The remaining $(N - N_f \times (K+1))$ sensors will be placed randomly and wouldn't be used. When a hazardous event occurs, nodes move along vertical axes for a distance $Dis = D/(K\sqrt{2})$. The following algorithm illustrates mobile nodes movement:

Algorithm 2: Sensor Nodes Movement

Input: N; D; R_c and *NList*: the list of sensor nodes
TS: the node movement time step in seconds.
Output: Monitored area sweeped

```
1:  N_f ←ceil( D /(2*sqrt(2)*R_c ))*2;
2:  K←floor( (N – N_f )/N_f )
3:  Dis←D/(K* sqrt(2) ); DSn← 0
4:  Sn← First(NList);
5:  For i from 1 to N_f do Sn ←Next( NList) End
6:  While (DSn < Dis) do
7:      While ( Sn) do
8:          Sn→move(Sn→x , Sn→y + V* TS, TS)
9:          Sn ←Next( NList)
10:     End
11:     DSn ← DSn + V* TS;
12: End
```

In the above algorithm, each sensor calls its own function *move(a, b, ts)* to go from its current position to the *(a, b)* position within *ts* time interval. The *move* function also updates node's position when the *(a, b)* are reached.

2) The circular strategy algorithms

Before detailing initial positions placement and movement algorithms of the circular strategy, we define:

- N_a : The number of sensors per axis.
- K : The number of axes.
- θ : The rotation angle defining the angular distance to be performed by each sensor node.
- V : The velocity of a sensor node at level 1 (the nearest level to the sink node). According to equation (2), nodes belonging to level i progress at a speed $V_i = V \times i$ since level i nodes are separated from sink node by a distance equal to $R_c \times i$.

Algorithm 3: Initial Sensor Nodes Positioning

Input: N; D; V ; R_c and (x_p, y_p)
Output: NList: Sensor nodes list

```
1:  N_a ←ceil(D/(2* R_c )); K←floor(N/ N_a )
2:  θ ← 2π /K; Sn→x ← x_p ; Sn→y ← y_p
3:  Add(Sn, NList);
4:  For i from 1 to K do
5:    For j from 1 to N_a do
6:      Sn ←new_Node();
7:      Sn→x ←sin((i-1)* θ )* R_c *j+ x_p
8:      Sn→y ←cos((i-1)* θ )* R_c *j+ y_p
9:      Sn→speed ← V *j
10:     Add(Sn, NList);
11:   End; End
```

When a hazardous event occurs, nodes move around the sink node until they go through an angular distance of θ. The following algorithm illustrates the sensor nodes movement without considering communication aspect between the nodes which will be detailed in Section 5. We use the *move* function introduced in algorithm 2.

Algorithm 4: Sensor Nodes Movement

Input: *NList*; N; R_c ; D; V and (x_p, y_p)
TS: the node movement time step in seconds.
Output: Monitored area sweeped

1: $N_a \leftarrow$floor(D/(2* R_c)); $K \leftarrow$ceil(N/ N_a)
2: $\theta \leftarrow 2\pi /K$; $\theta_f \leftarrow 0$; $\theta_1 \leftarrow V$ *TS/ R_c
3: **While** ($\theta_f < \theta$) **do**
4: $Sn \leftarrow$ First($NList$); $Sn \leftarrow$ Next ($NList$)
5: $\theta_f \leftarrow \theta_f + \theta_1$
6: **For** i **from** 1 **to** K **do**
7: **For** j **from** 1 **to** N_a **do**
8: $a \leftarrow$sin((i-1)* $\theta + \theta f$)* R_c *j+ x_p
9: b $\leftarrow$cos((i-1)* $\theta + \theta f$)* R_c *j+ y_p
10: $Sn \rightarrow move(a , b, TS)$
11: $Sn \leftarrow$ Next ($NList$)
12: **End**; **End**; **End**

IV. PEFORMANCE EVALUATION OF LIENAR AND CIRCULAR STRATEGIES

In this Section, we propose to evaluate the performance of the linear and circular strategies. Indeed, we evaluate the Monitored Area Sweep Time (MAST) which corresponds to the time required by the sensor nodes to scan all the monitored area coordinates to locate victims. Moreover, as in the circular strategy sensor nodes progress at different speeds on the different levels of a given axis, we also propose to evaluate the energy consumption of linear and circular strategies due to mobility. To evaluate the sensor nodes deployment strategies, we implement the algorithms 1 to 4 defined in Section 2 under the WSNet [3] simulator, an event driven simulator dedicated to wireless sensor networks. WSNet defines models written in C for the different network layers Simulation scripts can be configured through xml files. Using the WSNet-replay tool, we present in Figure 3 the initial node deployment screen using the circular strategy (The number of sensors N=40).

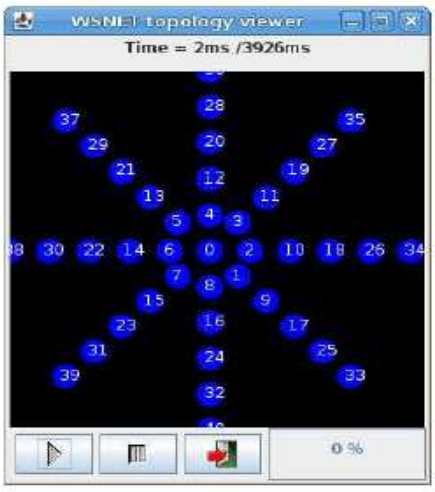


Figure 3. WSnet-replay screen for the circular strategy

In the next Section, we focus on area sweep time and energy consumption performance evaluation.

A. The Monitored Area Sweep Time

To derive results of the MAST parameter, we consider the communication range R_c equal to 30m, the velocity of mobile nodes in the linear strategy and level 1 nodes in the circular strategy equal to 2 m/s. In Figure 2, we fix the number N of sensor nodes deployed on the monitored area to 50 sensors, and depict the monitored zone sweep time as a function of the diagonal D of the monitored zone for both linear and circular strategies.

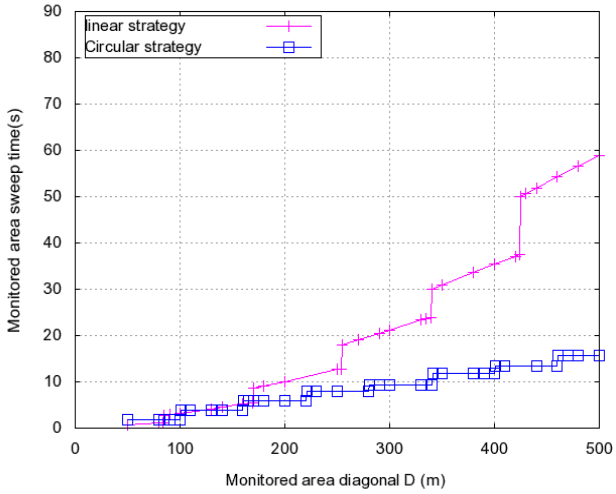


Figure 4. Monitoric Area Sweep Time

The curves show that with linear and circular strategies, we obtain step based sweep time values. This behavior can be explained by the use of *ceil()* and *floor()* functions. Moreover, curves show that the circular strategy sweep time growth is smoother than the one of the linear strategy as D increases. We can therefore conclude that it would be better to use the circular strategy when large areas are monitored. However, we can't conclude which strategy is better without considering the energy dissipation effect caused by nodes mobility.

B. Energy consumption due to mobility

In this paragraph, we focus on energy consumption caused by nodes mobility. Indeed, a pretty simple model of energy consumption due to mobility was introduced in [4] and states that the energy drops linearly with the distance (27.96 J/m).

To implement this energy dissipation model in simulation scenarios, we consider that each sensor node has an initial energy of 10^6 joules and we trigger a periodic event that computes the distance between the old node position and the current node position and evaluates the energy dissipation value. Figure 5 depicts the energy consumption of both linear and circular strategies as a function of D. Curves show that the energy consumed by the linear strategy is smaller than the one consumed by the circular strategy since nodes in circular

 ISBN:978-1-61208-113-7

strategy progress at different speeds. We can therefore conclude from sweep time and energy consumption values, that the circular sensors deployment strategy offers a better zone sweep time than the linear strategy at the expanse of high energy consumption.

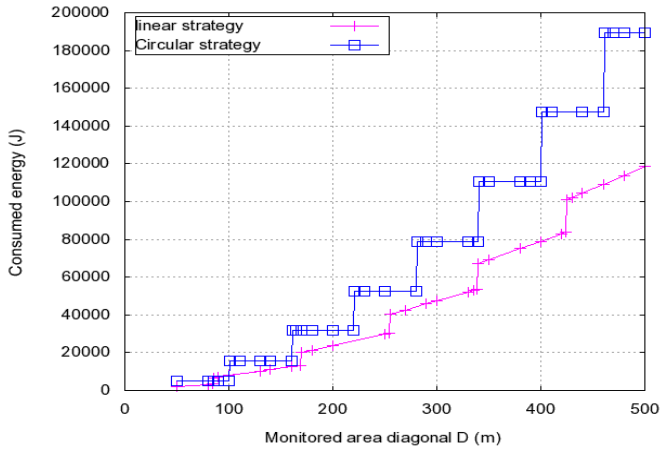


Figure 5. Energy consumption caused by sensors mobility

C. *Verification of simulation results*

In the above paragraphs, we depicted MAST and energy consumption values obtained by simulation. In this paragraph, we propose to verify the simulation results by considering analytical expressions of the monitored area sweep time and energy consumption due to mobility as a function of the number of sensors *N*, the monitored area diagonal *D*, the communication range R_c and the velocity *V*.

Indeed, in the linear strategy, as sensor nodes on horizontal axes progress simultaneously, the time required by sensor nodes to sweep all the monitored area is the time required by each mobile sensor on a horizontal axis to go through a $D/\left(K\sqrt{2}\right)$ distance. Therefore, the MAST parameter analytical expression is given by:

$$MAST = \frac{D}{\sqrt{2} * V * \left\lfloor N/2\left\lceil D/2\sqrt{2}/R_c \right\rceil - 1 \right\rfloor} \quad (3)$$

In the circular strategy, as nodes located on the same axis and on different levels have to progress simultaneously, the monitored area sweep time is defined by the time required by a node on level 1 to go through an angular distance of $\frac{2\pi}{K}$ with the velocity V, where K is the number of axes. Therefore:

$$MAST = \frac{2\pi R_c}{\left\lfloor N/\left\lceil D/2R_c \right\rceil \right\rfloor V} \quad (4)$$

where $\left\lfloor N/\left\lceil D/2R_c \right\rceil \right\rfloor$ is the analytical expression of the number of axes K. Using the same parameters values as for simulation scenarios, we depict in Figure 6 analytical and simulation results of the monitored zone sweep time as a function of the diagonal *D*. Curves show that for both strategies, analytical and simulation results coincide. In the same way, we can derive energy dissipation analytical expressions caused by nodes mobility.

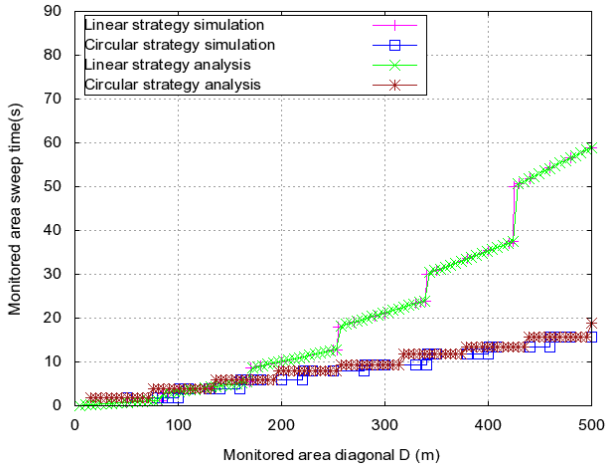


Figure 6. Analytical and simulation results of the monitored area sweep time

Indeed, in the absence of node communication, the energy decreases linearly with the distance (α joules/m). If the linear strategy is used, all the ($K * N_f$) mobile sensors, go through a $D/\left(K\sqrt{2}\right)$ distance. Therefore, the whole network consumption energy (WNCE) is:

$$WNCE = \frac{D * \alpha * K * N_f}{\sqrt{2} * K} = \frac{D\alpha}{\sqrt{2}} * 2\left\lceil D/2\sqrt{2}/R_c \right\rceil \quad (5)$$

When the circular strategy is considered, the energy consumed by level *i* sensors is *i* times the one consumed by level 1 sensors. Therefore:

$$WNCE = K\left(R_c\theta + 2R_c\theta + 3R_c\theta + \ldots N_a R_c\theta\right)\alpha$$
$$= 2\pi R_c \frac{N_a\left(N_a+1\right)\alpha}{2} = \pi R_c \alpha\left\lceil D/2R_c \right\rceil\left(\left\lceil D/2R_c \right\rceil + 1\right) \quad (6)$$

In Figure 7, we depict the energy consumption due to mobility for linear and circular strategies using both analysis and simulation results (the dissipation factor α= 27.96 J/m).

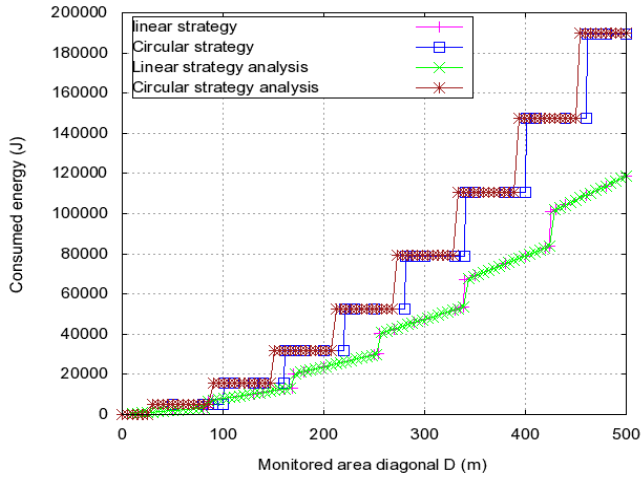


Figure 7. Analytical and simulation results of the energy consumption

As for the MAST parameter, analytical and simulations results of energy consumption due du mobility coincide. We can therefore conclude that sensor deployment strategies behaviors are correctly implemented under the WSNet simulator and this behavior can be extended to support nodes communications for victims' detection purposes.

V. VICTIMS DETECTION MECHANSIM

As the simulation modules implementation was verified analytically, we propose in this Section to extend the movement algorithms 2 and 4 detailed in Section 3 to support nodes communication so that mobile nodes can notify the sink node of victims' positions when they are detected.

Indeed, in the linear strategy, we propose a new mobility model that ensures communication between sensor nodes when a victim is detected. If a sensor node detects a victim, it sends a message to its neighbor belonging to the same horizontal axis and the nearest one to the fixed sensors. The receiving node will repeat the same procedure until the message reaches one of the vertical sensors that will forward the message to the sink node (*ID=0*). Therefore, the algorithm 2 related to the nodes movement in Section 3 becomes:

Algorithm 5: Sensor Nodes Victims Detection

Input: N; R_c ; *D, V; NList* and (x_p, y_p)
TS: the node movement time step in seconds.
Output: Victims positions detected.

1: $N_f \leftarrow$ceil($D/(2*sqrt(2)*R_c)$)*2;
2: $K \leftarrow$floor($(N - N_f)/N_f$); $Dis \leftarrow D/(K*sqrt(2))$;
3: $DSn \leftarrow 0$; $Sn \leftarrow$ First(*NList*); $Sn \leftarrow$ Next(*NList*)
4: **For** i **from** 1 **to** N_f **do** //Fixed sensors
5: *VicList* ←*Detect_victims*(*Sn*)
6: **if** (*VicList*) **then**
7: *msg*←*New_Msg( )*; Add(*msg*, *VicList*)
8: **End if**
9: **if** ($Sn \rightarrow y < y_p$) **then** *sg*←1 **else** *sg*← -1
10: **End if**
11: *Tx_victim*(*Sn*→*x*, *Sn*→*y*+ *D*/*Rc*/*sqrt(2)***sg* , msg) ; *Sn* ←Next*(NList)*
12: **End**
13: **While** *(DSn < Dis)* **do** //Mobile sensors
14: **While** *(Sn)* **do** *VicList* ←*Detect_victims*(*Sn*)
15: **if** (*VicList*) **then**
16: *msg*←*New_Msg( )*; Add(*msg*, *VicList*)
17: **End if**
18: **if** ($Sn \rightarrow x < x_p$) **then** *sg*←1 **else** *sg*← -1
19: **End if**
20: *Tx_victim*(*Sn*→*x* + *D*/ *Rc*/*sqrt(2)***sg* ,
21: *Sn*→*y*, msg) ;
22: *Sn*→*move(Sn*→*x* , *Sn*→*y* + *V** *TS, TS)*
23: *Sn* ←*Next(NList);*
24: **End**
25: *DSn* ← *DSn* +*V***TS*
26: **End**

In algorithm 5, the function *Detect_victims*(*Sn*) detects victims in *Sn* sensing range and returns their positions list. Then, *Sn* calls the recursive function *Tx_victim()* to transmit victims positions to its direct neighbor *Sn+1* on the same horizontal axis. As *Tx_victim()* is a recursive function, *Sn+1* will repeat the same procedure until the message reaches one of the vertical sensors that forwards the message to the sink node using the same function *Tx_victim()*.

In the circular strategy, we propose a new algorithm to ensure communication between sensor nodes. Indeed, when a mobile sensor detects a victim in its sensing range, it sends a message to the lower level node situated on the same axis. Sensor nodes on lower level axis will forward the message until it reaches the sink node (*ID=0*). Therefore, the algorithm 4 related to the nodes movement in Section 3 becomes:

Algorithm 6: Sensor Nodes Victims Detection

Input: N; R_c ; D; V ; *NList* and (x_p, y_p)
TS: the node movement time step in seconds.
Output: Victims positions detected.

1: $N_a \leftarrow$floor($D/2*R_c$); $K \leftarrow$ceil(N/N_a)
2: $\theta \leftarrow 2\pi/K$; $\theta_f \leftarrow 0$; $\theta_1 \leftarrow V*TS/R_c$
3: **While** ($\theta_f < \theta$) **do**
4: $Sn \leftarrow$ First(*NList*); $Sn \leftarrow$ Next (*NList*)
5: $\theta_f \leftarrow \theta_f + \theta_1$
6: **For** *i* **from** 1 **to** K **do**
7: **For** *j* **from** 1 **to** N_a **do**
8: *VicList* ←*Detect_victims*(*Sn*)
9: **if** (*VicList*) **then**
10: *msg*←*New_Msg( )*; Add(*msg*, *VicList*)
11: **End if**
12: $a \leftarrow \sin((i-1)*\theta + \theta_f)*R_c*j + x_p$
13: $b \leftarrow \cos((i-1)*\theta + \theta_f)*R_c*j + y_p$
14: $a1 \leftarrow \sin((i-1)*\theta + \theta_f - \theta_1)*R_c*j + x_p$
15: $b1 \leftarrow \cos((i-1)*\theta + \theta_f - \theta_1)*R_c*j + y_p$
16: *Tx_victim*((*a* – x_p)* (*a1* , *b1*, *msg*)
17: *Sn*→*move(a , b, TS);* $Sn \leftarrow$ Next (*NList*)
18: **End**;
19: **End; End**

In the same way, the recursive function *Tx_victim()* allows a sensor node *Sn* to forward victims positions to his direct neighbor on the same axis and the level below. Using the recursive *Tx_victim()* function victims' positions are forwarded to the sink node. In table 1, we compare the energy dissipation caused by mobility to the resulting energy dissipation caused by both mobility and communication:

TABLE I. ENERGY CONSUMPTION VALUES

Diagonal D (in m)	*Linear strategy*		*Circular strategy*	
	No commu-nication	*Communi-cation*	*No commu-nication*	*Communi-cation*
100	7902.62	7903.21	5266.56	5266.98
200	23707.87	23708.49	31599.39	31599.86
300	47415.75	47416.31	78998.48	78998.78
400	79026.25	79026.84	110597.88	110598.27
500	118539.38	118539.98	189596.31	189596.80

Table 1 shows that sensors nodes communication has a little impact on energy consumption and the major energy dissipation is due to mobility.

In Figure 8, we consider two simulation scenarios with different number of victims (20 and 80 victims). For these two scenarios, we depict the last victim detection time as a function of the monitored area diagonal *D*.

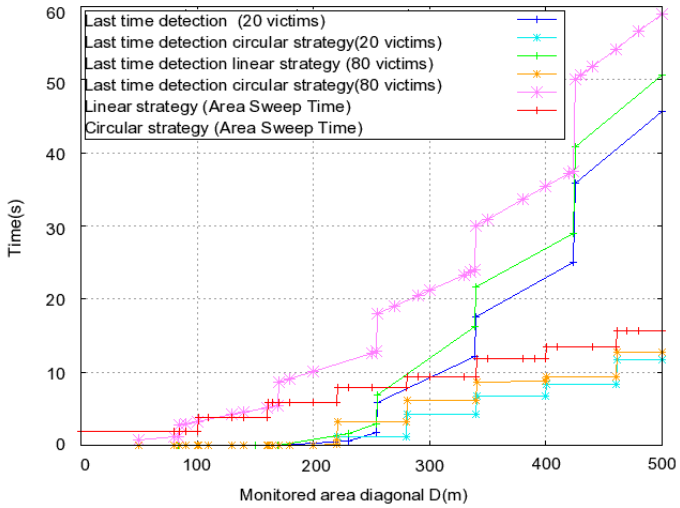


Figure 8. Comparision of last victim detection time and the MAST value

Figure 8 shows that the last victim detection time curves are under the MAST curve for both linear and circular strategies. Moreover, we notice that the last victim detection time curves become closer to the MAST curve as the number of victims increases. Therefore, we conclude that the area sweep time is an upper bound of the time required to detect all the victims within the monitored area and that the last victim detection time approaches the MAST value when the number of victims to be detected increases considerably.

VI. CONCLUSIONS AND FUTURE WORK

A. Conclusions

In this paper, we focused on the use of wireless sensor networks in rescue applications. This application consists in deploying a wireless sensor network in a monitored area to locate victims when a disaster occurs. Then, we considered mobile sensor nodes and focused on the way they are initially deployed and how do they move within the monitored area when a hazardous event happens. Therefore, we proposed two sensor deployment strategies: the linear strategy, where sensor nodes are arranged along horizontal and vertical axes and the circular strategy where sensors are arranged in a circular way around the sink node. For these two strategies, we proposed algorithms for initial nodes deployment and nodes movement within the network. Therefore, we implemented the algorithms of both strategies under the WSNet simulator. For These two techniques, we derived the MAST parameter which corresponds to the time required by the sensor network to scan all the monitored area coordinates to find victims. We also derived energy consumption due to mobility as mobile nodes in the circular strategy move at different speed values. Simulations results were easily verified using analytical expressions of the MAST and energy consumption values. The results showed that circular strategy is faster than linear strategy in terms of monitored area sweep time. However, the whole network energy consumption in the circular strategy is greater than the one of the linear strategy. In the last Section, we considered communication between mobile nodes and introduced message exchange mechanisms in both linear and circular strategies. We derived the energy consumption resulting in both nodes mobility and communication. We also verified by simulation that the MAST value is an upper bound of the time required detecting all the victims, which corresponds to the last victim detection time. Moreover, we noticed that the last victim detection time encloses the MAST value as the number of victims increases within the monitored area.

B. Future work

In our future work, we propose to evaluate the sensor nodes deployment strategies under more realistic conditions. Indeed, we'll consider that we are in presence of obstacles within the monitored area and we'll propose different solutions so that obstacles could be by bypassed by the sensor nodes. Moreover, when sensor nodes move within the monitored area, they either use GPS or other localization algorithms to determine their next positions within the monitored area. These localization techniques known to be approximate may lead sensor nodes to incorrectly determine their next positions within the monitored area (with an error factor *e*). Therefore, we'll evaluate the impact of on this incorrectness on the number of victims detected within the monitored area.

 ISBN:978-1-61208-113-7

REFERENCES

[1] A. Howard, M. J. Mataric, and G. S. Sukhatme, "An Incremental Self-Deployment Algorithm for Mobile Sensor Networks," *Autonomous Robots, Special Issue on Intelligent Embedded Systems*, September 2002, vol 13 (2), pp 113-126.

[2] B. Bhattacharya, M. Burmester, Y. Hu, E. Kranakis, and Q. Shi "Optimal Movement of Mobile Sensors for Barrier Coverage of a Planar Region", the 2nd international conference on Combinatorial Optimization and Application, 2008, pp. 5515-5528.

[3] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed Diffusion: A Scalable and Robust Communication," Mobicom 2000, pp 56-67.

[4] G. Wang, G. Cao, T. La Porta, and W. Zhang, "Sensor Relocation in Mobile Sensor Networks", Infocom 2005, vol. 4, pp. 2302- 2312.

[5] K. Sha, W. Shi, and O. Watkins, "Using Wireless Sensor Networks for Fire Rescue Applications: Requirements and Challenges", The 2006 IEEE International Conference on Electro/information Technology, pp. 239 -244.

[6] I. M. Rekleitis, A. P. New, and H. Choset, "Distributed coverage of unknown/unstructured environments by mobile sensor networks," 3rd International NRL Workshop on Multi-Robot Systems, 2005, pp. 145-155.

[7] M. A. Batalin and G. S. Sukhatme, "Multi-robot dynamic coverage of a planar bounded environment", Technical Report, CRES-03-011, 2003.

[8] N. Heo and PK Varshney. "A distributed self spreading algorithm for mobile wireless sensor networks. Wireless Communications and Networking, 2003, WCNC'2003, vol.3, pp. 1597-1602.

[9] N. Heo and P.K. Varshney, "Energy-efficient deployment of Intelligent Mobile sensor networks", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans , vol. 35(1), pp. 78-92, 2005.

[10] N. Xu, "A Survey of Sensor Network Applications", IEEE Communications Magazine, 2002, vol. 40(8), pp. 102-114.

[11] R. Severino and M. Alves, "Engineering a Search and Rescue Application with a Wireless Sensor Network - based Localization Mechanism", WoWMoM 2007, pp. 1- 4.

[12] S. C. Wong and B. A. MacDonald, "A topological coverage algorithm for mobile robots," in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003, pp. 1685-1690.

[13] W. R. Heinzelman, J. Kulik, and H. Balakrishnan, "Adaptive Protocols for Information Dissemination in Wireless Sensor Network," *ACM Mobicom*, August 1999, pp. 174 -185.

[14] http ://*wsnet.gforge.inria.fr*/ inria-00475581, version 1 - 22 Apr 2010.

[15] Y. Zou and K. Chakrabarty, "Sensor Deployment and Target Localization Based on Virtual Forces," *INFOCOM* 2003, vol.2, pp. 1293-1303.

 ISBN:978-1-61208-113-7

UHF-RFID-Based Localization Using Spread-Spectrum Signals

Andreas Loeffler
Chair of Information Technologies
Friedrich-Alexander-University of Erlangen-Nuremberg
Erlangen, Germany
Email: loeffler@like.eei.uni-erlangen.de

***Abstract*—Determining the range between RFID tags and readers, particularly in the UHF region, is subject of current research topics. This paper presents a simulation-based approach, using spread-spectrum techniques to reckon the range between UHF RFID tags and a reader system. The UHF-RFID reader reads out the surrounding tags within its reading range, with each tag containing information about its own location. Accordingly, the reader reads the location of each tag, hence being able to generate a basic map. To define the most likely position of the reader itself, it must estimate the distance to each of the RFID tags. This estimation is provided through a spread-spectrum approach utilizing the complete UHF bandwidth of approximately 150 MHz to achieve a high positioning resolution. This work proposes particular distance estimation techniques and presents results for such distance measurements.**

***Keywords*-Radiofrequency identification, Spread spectrum, Wideband, Communication channels.**

I. INTRODUCTION

The increase in sales of navigation systems [1] follows the trend to know exactly where objects or persons (mostly the person itself) are located. That is one of the reasons why global navigation satellite systems (GNSS) like GPS have become very popular. However, the usage of GNSS systems is rather limited to outdoor navigation. Multipath effects, high fading and blocking of satellite navigation signals usually limit the usage of satellite navigation systems in indoor areas.

In the past, indeed, great efforts were made to handle these drawbacks for the indoor area. WLAN-based systems, for instance, use the received signal strength (RSS) to determine the position of a WLAN device. Unfortunately, these systems need training sequences every now and then, resulting in setting up reference maps with reference points (also known as fingerprinting) [2]. However, WLAN-based navigation systems may be combined with inertial navigation systems to receive a much better accuracy. Ultra-wideband systems (UWB), unlike WLAN systems, calculate the *Time of Arrival* (TOA), *Time Difference of Arrival* (TDOA) and/or *Angle of Arrival* (AOA) in order to obtain the position. One example using UWB-technology is the Ubisense platform [3]. These UWB-based systems can achieve a high accuracy due to their very high bandwidth [4]. Besides the described indoor navigation systems exist a lot more techniques to determine the position, including Infrared, Bluetooth, GSM, etc. For instance, SpotON [5] and LANDMARC [6] are RFID-based navigation systems. Generally, both of these systems are based on measuring the RSS to get the current position. However, both systems, SpotON and LANDMARC, use active RFID tags resulting in higher costs per tag (manufacturing and service). Another RFID-based navigation system is described in [7] and [8]. This system is based on passive RFID transponders working at 13.56 MHz (RFID-HF) using the ISO14443 standard with the MIFARE extension. The major drawback of using RFID-HF technology is the low communication distance.

The work described in this paper pursues the following approach: The navigation system should depend on UHF-RFID technology to allow a higher distance to the tags. Also, the distance to be determined between reader and tag shall not depend upon fingerprints, reference maps, or any other training sequences, including RSS measurements. Nevertheless, the position of the reader shall be detected by evaluating an TOA approach from reader to tag (of course, several tags), whereas the, location-wise, fixed tags, contain their very own positions.

This paper shows some theoretical statements being proofed in different simulation scenarios, which deal with conditions mentioned above.

The paper is organized as follows. Section I gives a brief introduction, whereas Section II describes the scenario of the proposed work. The following section shows more details how the system itself is designed and which techniques are needed to achieve the position determination. Subsequently, Section IV is offering simulation results regarding various channel characteristics. Results are offered in Section V, followed by a conclusion and a reference to future work in Section VI.

II. SCENARIO

This section highlights the initial scenario the following sections will be build upon. Assuming an arbitrarily given room with usually UHF-based RFID transponders tagged to the wall(s) or other fixed, unmovable objects. Also, there is an RFID reader, equipped with an omni-directional antenna, somewhere in the room. Furthermore, the system is simplified by assuming that all antennas (reader and transponder)

 ISBN:978-1-61208-113-7

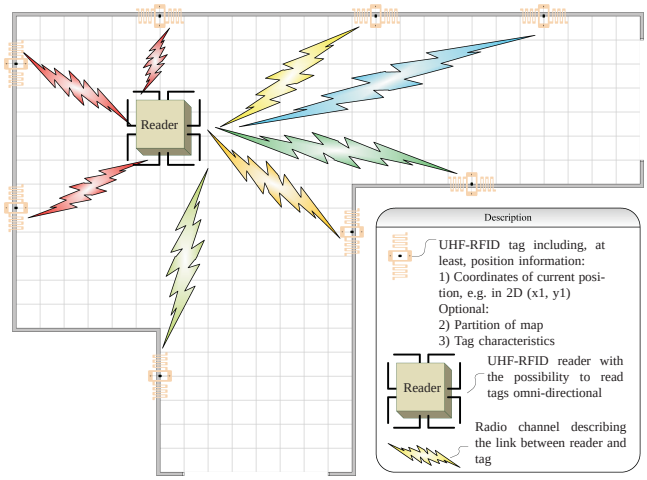


Figure 1. The underlying scenario to deal with

are positioned in the same plane (i.e., 2D, e.g., xy-plane). Figure 1 describes this scenario. The flashes between the reader and the transponders demonstrate the different radio channel links between each reader/transponder pair.

However, the distributed RFID transponders store position data. These position data, at least, consist of latitude, longitude and elevation or just relative coordinates. It is important that every transponder stores its own position; that means reading out a transponder would return its location within the environment. It is assumed that the RFID reader reads out all transponders within the reader's communication range. In that case, the reader knows actually where all the read out tags are located - so that the reader may build a virtual map including the location of every transponder. Unfortunately, the RFID reader itself does not know where it is located. Therefore, the task is to determine the position of the reader within the cloud of transponders, meaning to enable the reader to find its own position by evaluating the radio links to each of the transponders. Assuming additionally, that the reader is able to determine the distance to each transponder, the reader is able to locate itself by evaluating the positions of all transponders.

Of course, there are several ways of estimating the distance between RFID reader and RFID transponder. The previous Section I showed some methods how to estimate the current position using RFID technology, whereas the subsequent Section III describes the proposed approach to determine the distance between an RFID transponder and an RFID reader.

Eexemplary applications of such an RFID-based system for the indoor area would include pedestrian and automatic navigation as well as methods for item searching and position logging.

III. Proposed Work

For estimating the distance between an RFID transponder and the RFID reader, a method very much based on radar techniques [9], particularly pulse compression radar, is suggested. Of course, there are some major differences between a classical pulse compression radar approach and the following proposal.

The reader broadcasts, in addition to the carrier, a broadband signal in the range of the UHF transponders' bandwidth being between 100 MHz and 150 MHz. Due to the high bandwidth, the normalized *signal-to-noise ratio* S/N, also known as E_b/N_0 (energy per bit to noise power spectral density ratio), of the signal is very low. Wide bandwidth signals using low energy are, for instance, ultra-wideband (UWB) signals. UWB signals mainly occur in free-licensed frequency bands, as the signal power is mostly less than the average noise power. In Germany, since 2008, UWB signals below 1.6 GHz may have a spectral power density of -90 dBm/MHz (transmitted power). Assuming room temperature, the noise power per bandwidth is approximately -114 dBm/MHz at the receiver. A path loss of approximately 70 dB (describing the path loss among the way from the reader to the transponder and back) results in an S/N of -46 dB. Using such broad frequency bands and low S/N ratios limit the methods to determine the distance between tag and reader. That is the reason why spread-spectrum (SS) technique is introduced. First, to achieve a processing gain, and second, that due to the higher bandwidth a more accurate positioning resolution is more likely to achieve. The following subsections give some insights into the underlying techniques.

A. Signal Spreading

A technique for coping within such low-S/N environments is called spread-spectrum. Methods of this technique spread signals in order to achieve a higher bandwidth. This technique leads to less interference, jamming, undeliberate detection and some more features [10]. Of course, the transmitter's, but mainly the receiver's architecture gets more complex. Spread-spectrum technique is nowadays fairly common (e.g., WLAN, UMTS, etc.). The spreading method of choice in this work is direct-sequence spread spectrum (DSSS). This method takes the incoming (data) signal and multiplies it with a given spreading sequence. The sequence usually exists of several chips, which have a smaller time period than the bit rate. This leads to a spreaded signal comprising a higher bandwidth but with less signal power per Hertz. Therefore, the power of the signal is spreaded within the frequency domain. Despreading this spreaded signal at the receiver is realized by multiplying the incoming signal with the exact same spreading sequence the transmitter used. Doing so, transforms the former spreaded signal into a narrowband signal, again. This narrowband signal equals the

 ISBN:978-1-61208-113-7

unspreaded data signal in the transmitter, usually with some additive noise aspects in phase, frequency and amplitude.

1) Barker Codes: The performance of the distance determination depends very much on the implemented spreading sequences. There are a lot of sequences, which usually differ in autocorrelation performance, peak side-lobes, orthogonality, etc. As orthogonality is not of interest as there is no multi-user system (as e.g., described in [11]), only the performance of the autocorrelation and the distance to the highest peak side-lobes do matter. A good and simple choice of adequate codes are Barker codes [12]. Usually, Barker codes are used in radar and synchronization applications. The 13-Bit Barker code is the only one with a side-lobe of maximum $+1$. Another bunch of codes, called the PN-codes (pseudo noise codes), can be used, too. These codes may be created using shift registers. They have a slightly smaller performance compared to Barker codes, but with the advantage of being as long as wanted and, of course, flexible in means of being soft-coded (i.e. implemented in software). In order to keep it simple, the 13-Bit Barker code is used so far.

2) Cross-Correlation of Spreaded Signals: The question is how to get positioning data out of these spreaded data signals. As the transmitted signal is known *a priori*, and the reflected, received signal from the tag is known at the reader, too, both signals can be correlated to get the desired TOA measurement. Assuming a simple additive white Gaussian noise channel (AWGN), in which the received signal is only time-shifted compared to the transmitted signal (with more or less noise power); processing a cross-correlation between the transmitted signal (time delay = 0) and the received signal (time delay > 0), a correlation peak arises at the delayed moment of time. If the transmitted signal is shifted in time, that this correlation peak is at $\tau = 0$, then the time difference between transmitted and received signal may be evaluated, as well as the distance to the RFID tag. An example is given in Figures 2 and 3, in which the transmitted and the noisy received baseband signals (Figure 2, shifted by $t = 20$) are cross-correlated (Figure 3). The peak of the correlation function is found at $\tau = 19.88$. Due to the additional noise the peak of the signal is not directly at $\tau = 20$, as the peak detection is processed by using quadratic approximation.

B. Coherent Addition

In the example above, the S/N of the received signal is still 3 dB, leading to an appropriate cross-correlation. Unfortunately, the S/N ratio within a real system would be in the range of -50 dB and even less at the receiver. Therefore, the peak detection would fail very often respectively not delivering a proper result.

Improving the S/N of a signal could be achieved through the usage of coherent addition. This means, that the incoming signals are added to each other in such a manner, that the timing issues from on signal frame to another stays constant (coherent). In other words, the noise, which is assumed to be Gaussian distributed, is canceled out. According to the Cramer-Rao lower bound [13], the S/N ratio is linear proportional to the number of coherent additions, under the assumption that the error variance does not depend upon both parameters. This is proved in simulations and further discussed in Section IV.

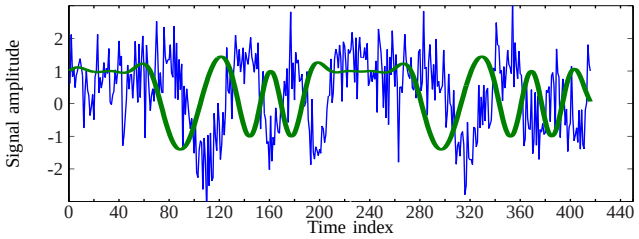


Figure 2. Transmitted (thick) and received (thin) baseband signal at the RFID reader (Inphase component)

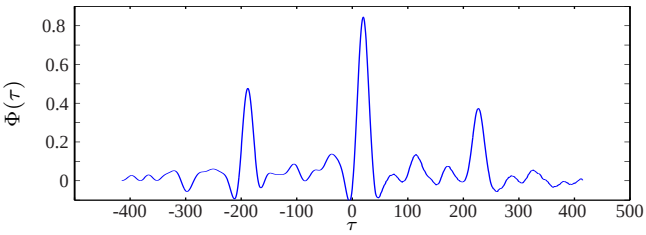


Figure 3. Cross-correlation $\Phi(\tau)$ between transmitted and received baseband signal (Figure 2) at the reader

To highlight the effect of coherent addition, Figure 4 shows the signal differences for an S/N of -20 dB with resulting signals taken at $n = 1$, 10, 100, 1,000 and 10,000 coherent additions. Easy to recognize, that the S/N is increasing with the number of additions.

The objective is to find an accurate (i.e., minimal) number of coherent additions n, to finally have a low error variance σ^2 for the peak detection, as this is a direct indicator for the distance measurement, because the peaks itself determine the distances between reader and tags.

C. Radio channels

As an example of the radio channels used in this work, two different frequency-selective fading channels (presence of multipath) are assumed. Both channels have an excess delay of $127 \times 100\,\text{ps} = 12.7\,\text{ns}$. The channels are build upon 128 discrete impulses, each with a distance of 100 ps to each other. The first pulse (at $\tau = 0$) is a Rician channel with Rician factor $K = 2$, whereas all the other channel impulses are Rayleigh distributed. The average path gain of each channel impulse is formed out of an exponential power delay profile $A_c[\tau]$ as in Equation (1).

$$A_c[\tau] = \frac{1}{T_0} \cdot e^{-\frac{\tau}{T_0}} \;,\; \tau = n \cdot 100\,\text{ps} \;\forall\; n \in [0;127] \cap \mathbb{N}_0 \quad (1)$$

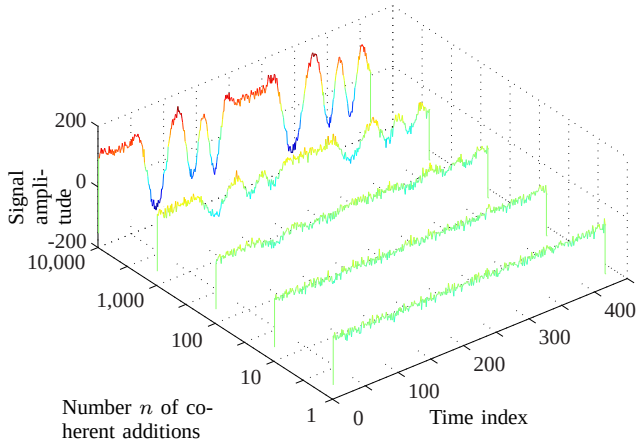


Figure 4. Coherent additions, increasing the S/N ratio

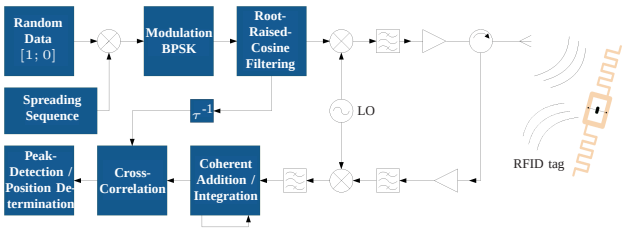
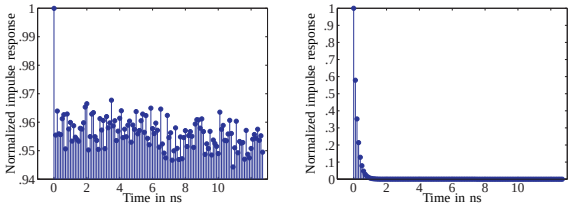


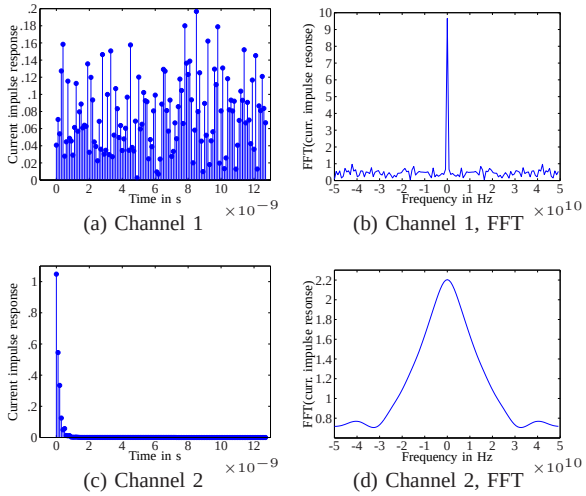
Figure 5. Reader architecture

By changing the value of T_0, different channel characteristics may be formed. A high value of T_0 compared to the symbol period T_s of the signal to be transmitted over the radio channel leads to a highly frequency-selective behavior, whereas a low value of T_0 will lead to a more *flat* channel characteristic. To underline this statement two different values are allocated to T_0 to describe two different radio channels. Channel 1 is characterized with $T_{0,1} = 100 \cdot T_s$, Channel 2 with $T_{0,2} = 0.01 \cdot T_s$. Figure 6 shows the *average* impulse responses of both channels. *Average* means, that the average impulse response consists of the mean of 10.000 simulated channel characteristics as every sub channel impulse is either Rician (at $\tau = 0$) or Rayleigh (at $\tau \neq 0$) distributed. The frequency characteristic of both channels is shown in Figure 7b and 7d together with a one



(a) Channel 1 with $T_{0,1} = 100 \cdot T_s$ (b) Channel 2 with $T_{0,2} = 0.01 \cdot T_s$

Figure 6. Average impulse responses of Channel 1 and Channel 2



(a) Channel 1 (b) Channel 1, FFT

(c) Channel 2 (d) Channel 2, FFT

Figure 7. Current impulse responses (time and frequency) of Channel 1 and Channel 2

channel time domain representation. As expected, Channel 1 shows a lower coherence bandwidth than Channel 2.

IV. Simulation of Various Scenarios

This section deals with simulations carried out, to evaluate the system's performance and limitations for given radio channels. The first subsection will present simulations for AWGN channels, whereas the following subsections will refer to more realistic radio channels, including real flat fading channels (no multipaths) and frequency-selective radio channels (number of multipaths > 0). All simulations were carried out using a symbol period of $T_s = 10\,\text{ns}$ by using BPSK (binary phase shift keying) and an RRC (root raised cosine)-filter with a roll-off factor of $\beta = 0.5$. This leads to an occupied bandwidth of about 150 MHz. Furthermore, all simulations use the 13-Bit Barker code, which was applied to spread 10 bits of data. The simulations are based on the model as given in Figure 5.

The simulation results show the error variance of the detected peaks σ_p^2 over the number of coherent additions n, whereas the variance refers to the variation of the peaks around its mean value. The upper theoretical limit $\sigma^2_{p,\text{theor.}}$ for the error variance, also referred to as $S/N = \text{-}\infty\,\text{dB}$, is calculated as in Equation (2). The search room for the peaks is arbitrarily limited to about $\pm$ one chip (exact: $1\frac{1}{16}$) with a sample rate of 16 bit per chip. This leads to the limitation of ± 17 in Equation (2).

$$\begin{aligned} \sigma^2_{p,\text{theor.}} &= 1/12\,(\text{Limit}_{\text{upper}} - \text{Limit}_{\text{lower}})^2 \\ &= 1/12\,(2 \cdot 17)^2 = 96.\bar{3} \end{aligned} \quad (2)$$

The theoretical limit is calculated by assuming a continuous uniform distribution of the peaks within the search room and matches an S/N of $\text{-}\infty\,\text{dB}$. The number of channel simulation repetitions in order to receive a significant variance for a given n was set to 100.

 ISBN:978-1-61208-113-7

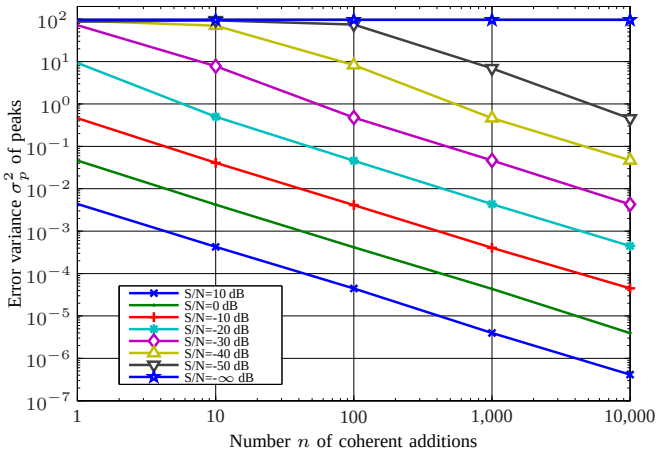


Figure 8. Simulation results for AWGN channel

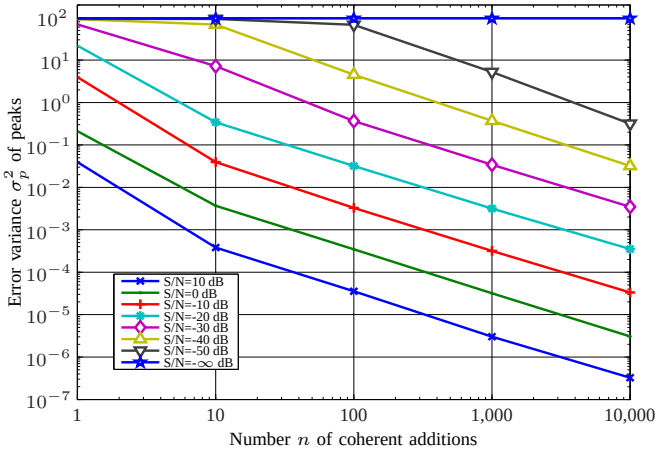


Figure 9. Simulation results for flat fading channel

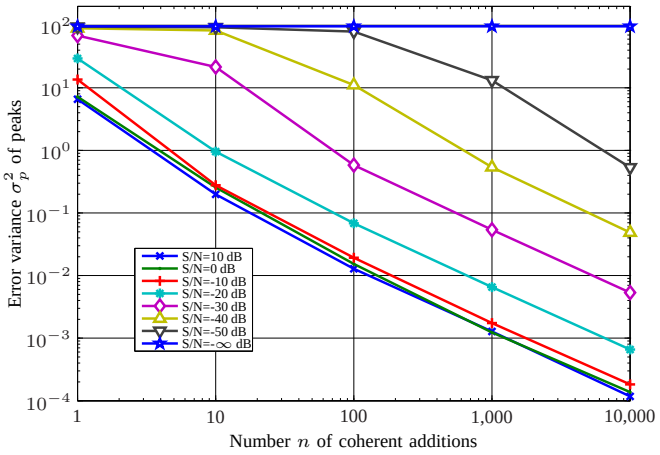


Figure 10. Simulation results for frequency-selective fading channel with $T_0 = 100\,\text{ps}$

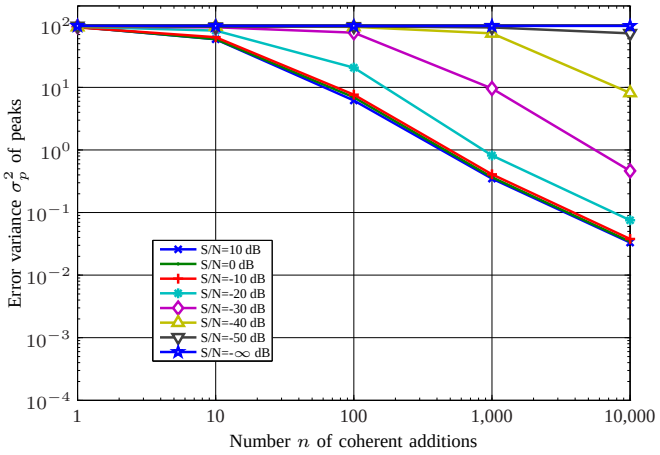


Figure 11. Simulation results for frequency-selective fading channel with $T_0 = 1\,\mu\text{s}$

A. Simulation for AWGN Channel

For a given S/N respectively E_b/N_0 the variance of the error is described over the number of coherent additions n. Refer to Section III for the details of the process.

Figure 8 shows the error variance σ_p^2 over the number of coherent additions n.

B. Simulation for Flat Fading Channel

This subsection shows the simulation results of a flat fading channel, i.e. a Rician distributed radio channel characteristic with no multipaths. Figure 9 shows the simulation results. The K-factor of the Rician channel was $K = 2$.

C. Simulation for Frequency-Selective Fading Channels

This subsection shows the simulation results of two frequency-selective fading channels, inheriting a Rician distributed radio channel characteristic with 127 Rayleigh distributed multipaths. The first simulation (Figure 10) refers to the frequency-selective fading Channel 2 (Figure 6b) with an $T_{0,2} = 0.01 \cdot T_s = 100\,\text{ps}$. As the signal symbol period $T_s = 10\,\text{ns}$ is greater (by a factor of 100) compared to $T_{0,2}$, this channel can be described as rather flat. On the other hand, the frequency-selective fading Channel 1 (Figure 6a) with an $T_{0,1} = 100 \cdot T_s = 1\,\mu\text{s}$ is really frequency-selective as the symbol period is much smaller (factor 100) than $T_{0,1}$. The simulation results of Channel 1 are shown in Figure 11.

V. Results

The simulations carried out in Section IV show three aspects. One is the fact, that the theoretical assumption of the error variance σ_p^2 being linear dependent on the reciprocal of the S/N and the number of coherent additions n could be proofed through simulation. This conclusion is valid for the AWGN channel and for the flat fading channel, at least if the values of the error variance are not close to

 ISBN:978-1-61208-113-7

the theoretical limit. The same assumption is almost valid for the frequency-selective channel, on condition, that the symbol period is much longer than the overall (average) channel excess time, which, of course, leads to a more flat fading channel characteristic. As expected, the error variance increases for a *real* frequency-selective channel (symbol period is smaller than the overall (average) channel excess time), as seen in Figure 11.

The second aspect is the upper theoretical limit of the error variance, which is valid for all channel simulations; refer to Figures 8, 9, 10 and 11.

The third aspect concerns the lower limit of the simulations regarding the frequency-selective fading channels (Figure 10 and 11). An improvement of the S/N to values greater than -10 dB would not lead directly to a smaller error variance in contrast to the AWGN channel (Figure 8) and the flat fading channel (Figure 9). This phenomenon may be explained through the influence of the intersymbol interference (ISI) of the described frequency-selective channels.

Furthermore, an important issue to focus on is the choice of an appropriate channel model. This model may be generated by either measuring out indoor areas (site-specific approach) or defining a statistical model (site-general approach). Both approaches show advantages, but for an ubiquitous estimation, at least for indoor environments, the site-general approach is more useful (see Equation (1)). This may easily understood by assuming that indoor environments change very fast (doors opening and closing, people, interior positions change, etc.). Anyway, a statistical model should be close to a worst-case scenario to provide a lower bound for the error variance of the distance.

Last carried out simulations show, that with a given maximum standard deviation σ_x of 15 cm, a sample rate of $f_s = 1.6$ GHz, and an occupied bandwidth of 150 MHz, the minimum number of coherent additions, to achieve that standard deviation, is 1877 . This result is valid for an AWGN channel model with an average S/N of -40 dB. Finally, the total measurement time is approximately 2.44 ms.

VI. Conclusion & Future Work

The paper presents an approach to estimate the position of an RFID reader by detecting the distance to the surrounding UHF-RFID transponders. As the tags contain their very own location, the reader may generate a basic map with the transponders' locations and its own position by evaluating the distances to the tags with the usage of various techniques. These techniques include spread-spectrum, coherent addition and cross-correlation approaches. The reader exploits the maximum bandwidth the UHF-tags provide (approximately 150 MHz) to achieve a high positioning resolution. In order to estimate the error variance σ_p^2 of the distance to the tags, various simulations were carried out considering varying channel characteristics to get a more realistic valuation of indoor attributes.

Future work would include, more channel characteristics as well as the non-linear behavior of the underlying UHF-RFID tags.

References

[1] RNCOS E-Services Private Limited, "World GPS Market Forecast to 2013," *Research and Markets*, Apr 2009. [Online]. Available: http://www.researchandmarkets.com/reportinfo.asp?report_id=836704

[2] P. Bahl and V. Padmanabhan, "RADAR: An in-building rf-based user location and tracking system," in *IEEE infocom*, vol. 2. Citeseer, 2000, pp. 775–784.

[3] P. Steggles and S. Gschwind, "The Ubisense smart space platform," in *Adjunct Proceedings of the Third International Conference on Pervasive Computing*, vol. 191, 2005.

[4] B. Waldmann, R. Weigel, and P. Gulden, "Method for high precision local positioning radar using an ultra wideband technique," in *Microwave Symposium Digest, 2008 IEEE MTT-S International*, June 2008, pp. 117–120.

[5] J. Hightower, R. Want, and G. Borriello, "SpotON: An indoor 3D location sensing technology based on RF signal strength," *UW CSE 00-02-02, University of Washington, Department of Computer Science and Engineering, Seattle, WA*, 2000.

[6] L. Ni, Y. Liu, Y. Lau, and A. Patil, "LANDMARC: indoor location sensing using active RFID," *Wireless Networks*, vol. 10, no. 6, pp. 701–710, 2004.

[7] A. Loeffler, U. Wissendheit, H. Gerhaeuser, and D. Kuznetsova, "GIDS - A system for combining RFID-based site information and web-based data for virtually displaying the location on handheld devices," in *RFID, 2008 IEEE International Conference on*, April 2008, pp. 312–319.

[8] A. Loeffler, U. Wissendheit, and D. Kuznetsova, "Using RFID-Capable Cell Phones for Creating an Extended Navigation Assistance," in *IMOC 2009*, nov 2009.

[9] J. Taylor, *Ultra-wideband radar technology*. CRC Press, 2001.

[10] M. Simon, J. Omura, R. Scholtz, and B. Levitt, *Spread spectrum communications handbook*. McGraw-Hill New York, 1994.

[11] A. Loeffler, F. Schuh, and H. Gerhaeuser, "Realization of a CDMA-Based RFID System Using a Semi-Active UHF Transponder," in *Wireless and Mobile Communications, 2010. ICWMC 2010. Sixth International Conference on*, September 2010, pp. 5–10.

[12] R. Barker, "Group synchronizing of binary digital systems," *Communication theory*, pp. 273–287, 1953.

[13] G. MacGougan, K. O'Keefe, and R. Klukas, "Ultra-wideband ranging precision and accuracy," *Measurement Science and Technology*, vol. 20, no. 9, p. 5105, 2009.

 ISBN:978-1-61208-113-7

An Error Reduction Algorithm for Position Estimation Systems Using Transmitted Directivity Information

Hiroyuki HATANO
Faculty of Engineering,
Shizuoka University
3-5-1 Johoku, Naka-ku,
Hamamatsu-shi, Shizuoka
432-8561, JAPAN
thhatan@ipc.shizuoka.ac.jp

Tomoharu MIZUTANI
Graduate School of Engineering,
Shizuoka University,
3-5-1 Johoku, Naka-ku,
Hamamatsu-shi, Shizuoka
432-8561, JAPAN
f0930142@ipc.shizuoka.ac.jp

Yoshihiko KUWAHARA
Faculty of Engineering,
Shizuoka University
3-5-1, Johoku, Naka-ku,
Hamamatsu-shi, Shizuoka
432-8561, JAPAN
tykuwab@ipc.shizuoka.ac.jp

***Abstract*—We consider a position estimation system for targets that exist in near wide area. The system has multiple sensors and estimates position using multiple receivers. Previously, receivers arranged in a straight line would generate a large position error in the direction of the line. In order to reduce this error, we herein propose a novel estimation algorithm using the directivity information of the transmitter. The proposed system uses the directional emission of an array of antennas in a transmitter. In the present paper, the error characteristic to be solved is introduced. The proposed algorithm is then presented. Finally, the error reduction performance is demonstrated through computer simulations. The obtained results indicate good error reduction performance.**

***Keywords*-sensor network; position estimation; localization; radar network; directivity**

I. INTRODUCTION

Interest in position estimation systems has been growing. In this paper, we focus on the estimation of the position of targets in the near wide area. Our position estimation systems are built with multiple sensors that are connected with networks (Figure 1). These sensors achieve reliable detection and accurate position estimation. Good performance can be achieved using even inexpensive devices. Moreover, networked sensors can cover a wide detection area. Several attractive applications of position estimation systems have been suggested, including indoor monitoring systems and near-range automotive radars [1], [2].

We assume that the sensors in the network can output only measured ranges (a measured range list) to the targets because of low cost and the simplicity of the components used to construct the sensors. The estimator must calculate target positions with high accuracy from only measured range lists provided by multiple sensors. For accurate positioning, it is important to discuss data processing of position estimation, which deals with measured range data from all of the sensors.

Over the past few years, several algorithms have been developed for position estimation using multiple sensors. As related work with multiple sensing devices, several studies have been made on sensor network systems [3]–[7]. Radar network systems have also been proposed [8]–[10]. Based on previous studies, trilateration techniques using geometric operations are popular for estimating target positions. These techniques are not optimum in terms of accuracy with respect to position estimation and may also detect "ghost targets", which are falsely detected about non-existent targets. This often occurs when the measured errors are large [1], [10].

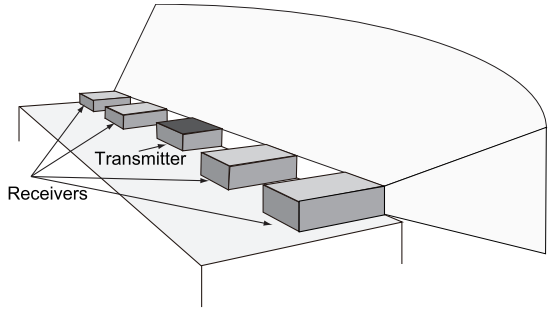


Figure 1. Position estimation system

In other techniques, measured ranges are treated as stochastic variables [11]–[13]. Typical techniques include estimation using minimum mean square error (MMSE) and maximum a posterior probability (MAP). The accuracy of these techniques is high compared to basic trilateration techniques. However, applying the MMSE method to the estimation of multiple targets is not easy. Among stochastic methods, the accuracy of the MAP method is optimum. However, data processing in the MAP method is complex. Therefore, we proposed a novel estimation algorithm, namely, the existence probability estimation method (EPEM). The EPEM calculates the existence probability of targets and estimates the target positions [14]. In the proposed method, the measured ranges are also treated as stochastic values. Moreover, the proposed method has approximately the same estimation accuracy and a lower

 ISBN:978-1-61208-113-7

calculation burden compared to MAP methods.

However, the estimation errors depend on the layout of the receivers. In particular, for the case in which the receivers are arranged in a straight line, large errors are generated in the direction of the line. Usually, a straight-line arrangement is useful because such an arrangement is easy to build and can be set up within a limited space. Thus, a technique to reduce the estimation errors is needed.

The goals of the present paper are as follows:

- Introduction of the EPEM algorithm,
- Clarification of the error performance depending on the sensor arrangement and description of the problem,
- Proposal of the existence probability estimation method with directivity information (EPEMD) algorithm,
- Evaluation of the error reduction through computer simulations.

The existence probability of the EPEMD is calculated using the directivity information of a cooperative transmitter. In the case of radar network systems, transmitters often use a directivity scan in order to reduce misdetections and expand detectable ranges for a limited power [15]. In the case of sensor network systems, the construction of electrical directivity antennas is advantageous because the sensor device requires a long deliverable range with low power. Therefore, it is meaningful to propose an estimation algorithm that considers the directivity pattern. As such, we evaluate the error reduction through computer simulations.

The remainder of this paper is organized as follows. In Section II, we present the system model and assumptions of the present study. In Section III, we introduce the EPEM algorithm, which is a position estimation algorithm. The algorithm is explained analytically, and the error performance and problems are presented. In Section IV, the proposed EPEMD algorithm is presented. In Section V, the performance of the error reduction is evaluated. Finally, Section VI summarizes the present study and presents suggestions for further research.

II. System Model of the Position Estimation System

We consider an estimation system with a transmitter and multiple receivers (Figure 1). Figure 2 and 3 show the system model. Figure 2 shows the sensor layout and the targets for position estimation. The numbers of receivers and targets are K and N, respectively. The origin of the coordinate system is the center of the receivers. The target position is given as (x_n, y_n), $1 \leq n \leq N$. Each receiver is assumed to be located on the x-axis. The x positions of the receivers are $\alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_K$.

The kth receiver outputs a measured range list composed of the ranges, namely, $\tilde{R}_k = (\tilde{r}_{k1}, \tilde{r}_{k2}, ..., \tilde{r}_{kN_k})$. Here, $N_k (\leq N)$ is the number of ranges included in the measured range list $\tilde{R}_k$. We assume the existence of a only direct path between target and transmitter/receiver. Subscript $(\tilde{\ })$ indicates measured values.

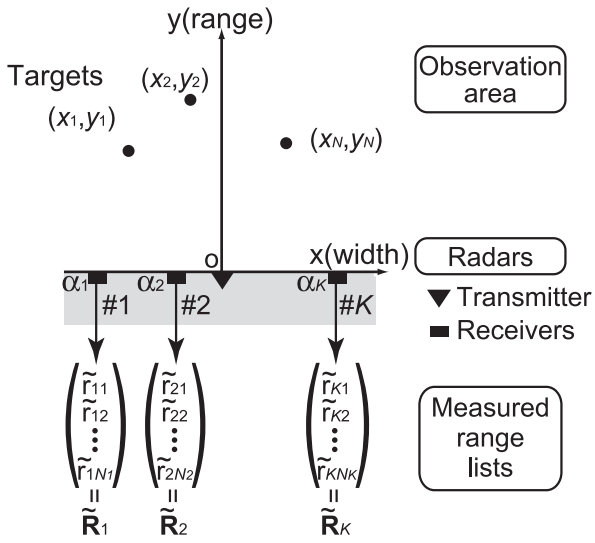


Figure 2. Layout of sensors and targets

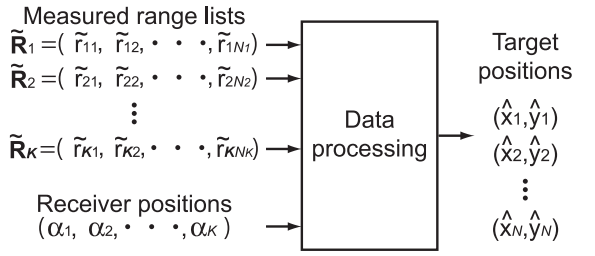


Figure 3. A data flow

Each measured range $\tilde{r}_{kn}$ in the list includes a measurement error:

$$\tilde{r}_{kn} = r_{kn} + \epsilon_k, \tag{1}$$

where r_{kn} is the true range between the nth target and the transmitter / the kth receiver, and ϵ_k is a stochastic variable, the variance of which is denoted as σ_k^2. Using the measured range lists obtained from all receivers and the position of the receivers, the target positions are estimated as shown in Figure 3.

III. Estimation Algorithm

In this section, we first introduce the proposed position estimation algorithm using the existence probability. The estimation characteristics are then summarized, and the problem is described.

A. Existence probability estimation method (EPEM)

Commonly used trilateration methods have a number of problems, as mentioned in Section I. In order to address these problems, the proposed estimation method, which is described below, deals with the measured rages as stochastic variables.

In order to estimate the target positions using the measured range lists provided by the receivers, we consider the following existence probability, which includes the conditional probability:

$$P(\hat{x}, \hat{y}|\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \cdots, \tilde{R}_K). \quad (2)$$

The probability of Equation (2) is the conditional probability of the target existence at $(\hat{x}, \hat{y})$ when the measured range lists $\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \cdots, \tilde{R}_K$ are obtained. Using Bayes' theorem, Equation (2) may be written as follows:

$$\frac{P(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \cdots, \tilde{R}_K|\hat{x}, \hat{y})}{P(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \cdots, \tilde{R}_K)} \cdot P(\hat{x}, \hat{y}). \quad (3)$$

In Equation (3), the denominator does not depend on the estimated parameter $(\hat{x}, \hat{y})$. Therefore, when $P(\hat{x}, \hat{y})$ is distributed uniformly, Equation (3) may have the same distribution shape:

$$P(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \cdots, \tilde{R}_K|\hat{x}, \hat{y}). \quad (4)$$

Each receiver is independent. Equation (4) may be expressed as follows because the measured range is an independent Gaussian variable:

$$\prod_{k=1}^{K} P(\tilde{R}_k|\hat{x}, \hat{y}). \quad (5)$$

Considering the combinations of targets and ranges, Equation (5) may be expressed as follows:

$$\prod_{k=1}^{K} \sum_{n=1}^{N} P(\tilde{r}_{kn}|\hat{x}, \hat{y}) \quad (6)$$

where we assume no pre-knowledge of the targets. In other words, the receivers do not know the relationship between the targets and the measured ranges. Next, the estimated parameters $(\hat{x}, \hat{y})$ can be transformed with the range from the transmitter/kth receiver to the target, i.e., $\hat{r}_{kn}$ as follows:

$$\hat{r}_{kn} = \sqrt{(\hat{x} - \alpha_k)^2 + \hat{y}^2} + \hat{x}^2 + \hat{y}^2 \quad (\text{for all } n). \quad (7)$$

Using the above relationship, Equation (6) is transformed into the following:

$$\prod_{k=1}^{K} \sum_{n=1}^{N} P(\tilde{r}_{kn}|\hat{r}_k). \quad (8)$$

The probability of $P(\tilde{r}|\hat{r})$ indicates the error characteristic of the known receiver. Using Equations (7) and (8), the distribution of the probability of the target existence at position $(\hat{x}, \hat{y})$ can be calculated when the measured ranges $\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \cdots, \tilde{R}_K$ are obtained. By selecting the local maximums of the distribution of Equation (8), the target positions can be estimated. The EPEM has approximately the same estimation accuracy as the MAP method, which is optimum in terms of maximum a posteriori probability [14].

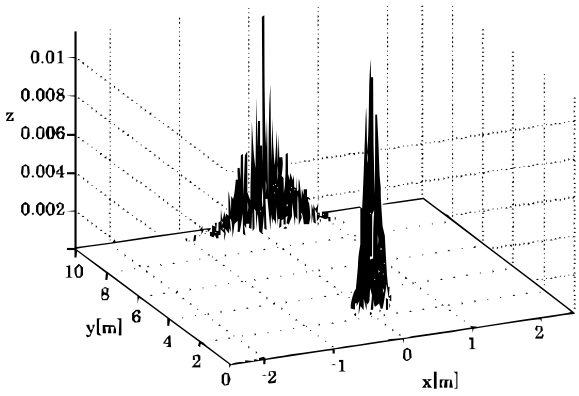


Figure 4. Distribution of estimated target positions (EPEM)

Table I
SIMULATION PARAMETERS 1

Number of receivers: K	3
Number of targets: N	2
Target positions	#1 $(x_1, y_1) = (0, 2)[m]$ #2 $(x_2, y_2) = (0, 9)[m]$
Array width of receivers	2 m
Distribution of measurement error $\tilde{r}$	Gaussian distribution ($\sigma_k = 0.075$)
Number of iterations	20,000

B. Estimation characteristics and problems

In the following, we present the estimation characteristics of the EPEM algorithm described in the previous section. The simulation parameters are shown in Table I. In the simulations, we assume the measurement error to be 0.3 m, which is typical [16]. According to this value, we set the standard derivation σ_k of the measured ranges ($4\sigma_k = 0.3$ [m]). The estimation trials of the targets are simulated. The trials generate the distribution of estimated positions. The results are shown in Figure 4. Moreover, the mean and variance of the distribution for each target position in Figure 4 are summarized in Table II. Figure 4 and Table II show that the error in the x-direction is larger than that in the y-direction. The reason for this is that the receivers are arranged along the x-axis. That is, large errors are generated in the direction of the receivers. In order to reduce the x-axis errors, we propose an estimation algorithm that uses the directivity of the transmitter.

IV. ESTIMATION ALGORITHM USING THE DIRECTIVITY OF THE TRANSMITTER

In this section, we first propose the estimation algorithm to solve the problem of the large error described in the pre-

Table II
CHARACTERISTICS OF THE ESTIMATED TARGETS (EPEM)

	Target 1	Target 2
E[$\hat{x}$] [m]	-0.001	-0.022
Var[$\hat{x}$]	0.014	0.240
E[$\hat{y}$] [m]	1.997	8.987
Var[$\hat{y}$]	0.002	0.002

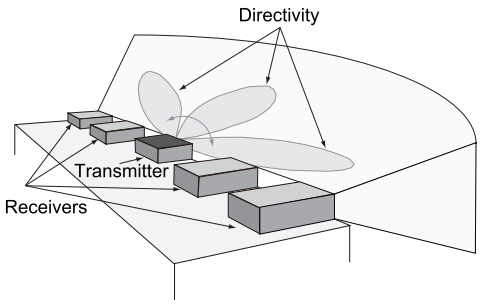


Figure 5. Image of the proposed system

vious section. The proposed algorithm is named as EPEMD (the existence probability estimation method with directivity information). Figure 5 shows an image of the proposed system. The difference between Figure 1 and Figure 5 is that the transmitter of Figure 5 has directivity.

A signal is radiated from the transmitter, which is composed of two or more antennas to achieve directivity. The reflected signal from the target is received by the receivers, which are placed along a straight line (x-axis). We introduce a transmission array antenna composed of L antennas, as shown in Figure 6. The antennas, which are centered at the origin of the coordinate system, are arranged symmetrically. The positions of the transmission antennas are assumed to be $(\beta_1, 0), (\beta_2, 0), \ldots, (\beta_L, 0)$. The variables A_l and $s_l(t)$ indicate the amplitude coefficient and the radiated signal from the lth antenna, respectively. The total signal in the θ direction is as follows:

$$S_{\text{sum}}(\theta, t) = s(\theta, t) \sum_{l=1}^{L} A_l \exp\{j2\pi f_0 (\frac{\beta_l}{c} \sin\theta)\} \quad (9)$$

where f_0 is the center frequency of the signal, and c is the speed of light. The based signal and common characteristic of the antennas, such as the directivity pattern of the element, is substituted as $s(\theta, t)$. In the present study, $|S_{\text{sum}}(\theta, t)|$, which indicates the gain generated by the array, is named as the directivity response pattern.

We attempt to reduce the estimated errors using this directivity response pattern. An example of a directivity response pattern is shown in Figure 7. Given this directivity response pattern, the signal can be reflected only from obstacles that exist in the area within the beam, such as Target #1. In contrast, Target #2 does not reflect the signal.

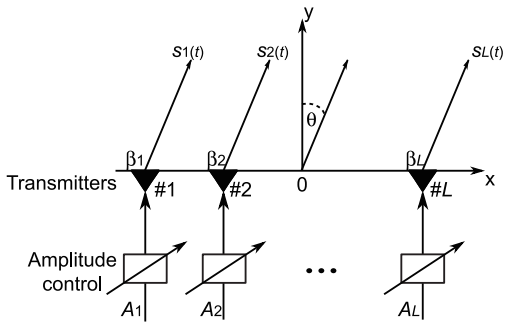


Figure 6. Structure of the transmitter

The function to specify the detectable area is as follows:

$$D_p(x, y) = \begin{cases} 1 & \text{(area that can be detected)} \\ 0 & \text{(area that cannot be detected)} \end{cases}. \quad (10)$$

Equation (10) expresses the area in which the target can reflect the signals or not. The reflectable area of the x-y plane can be calculated from the directivity response pattern.

The directivity response pattern is converted to the reflectable area of the x-y plane using the following radar equation:

$$S = \frac{\gamma P_t}{R^4} \quad (11)$$

where S is the electric power of the reflected signal, i.e., the signal received at the receiver, and γ is determined on the basis of, for example, the antenna gain and the effective reflection area of the targets. In addition, P_t is the power of the transmitter, and R is the range from the transmitter/receivers to the target. If S is defined as the minimum power detectable at the receiver, then the R is the maximum detectable range. Equation (11) can be rewritten as follows:

$$R = \sqrt[4]{\frac{\gamma}{S}} \sqrt[4]{P_t} \quad (12)$$

Next, we assume that the transmitting power becomes δP_t, that is δ times. Then, maximum detectable range R' can be rewritten in terms of R as follows:

$$R' = \sqrt[4]{\frac{\gamma}{S}} \sqrt[4]{\delta P_t} = \sqrt[4]{\delta} R \quad (13)$$

As mentioned above, $|S_{\text{sum}}(\theta, t)|$ in Equation (9) is the gain of the array, which is related to δ. When the gain of the electric power is $|S_{sum}(\theta, t)|$, the maximum detectable range R' can be calculated from Equations (9) and (13).

As a result, from Equations (10) and (8), we obtain the following equation:

$$\left[\prod_{k=1}^{K} \sum_{n=1}^{N} P(\tilde{r}_{kn} | \hat{r}_k)\right] \cdot D_p(x, y) \quad (14)$$

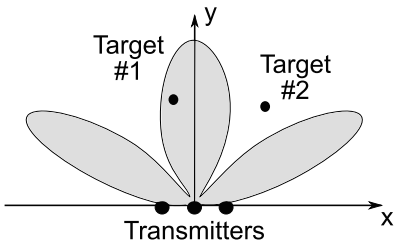


Figure 7. Directivity response pattern and targets

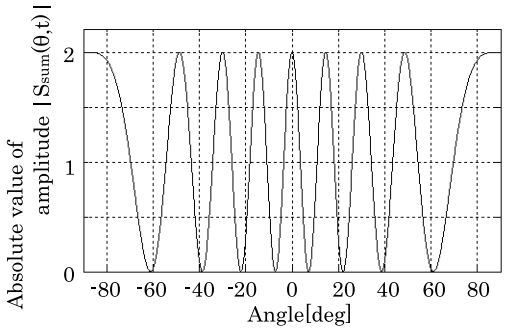


Figure 8. Designed directivity response pattern

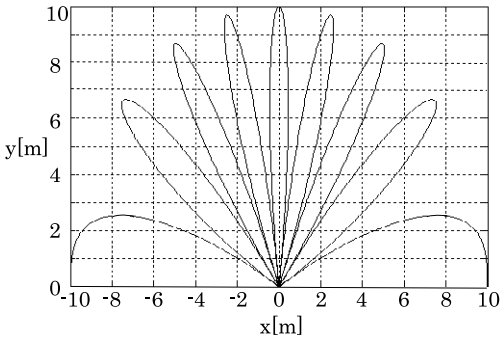


Figure 9. Detectable area

Table III
PARAMETERS OF THE TRANSMITTER

Frequency: f_0	24 [GHz]
Number of transmitters: L	3
Width of array [m]	0.1
Element positions [m]: B_l	-0.05, 0, 0.05
Amplitude control: A_l	0.5, 1, 0.5

Equation (14) gives the existence probability considering the directivity of the transmission signal. The generation of the directivity often results in null directions. In order to compensate for the null directions, the direction of the main lobe of the directivity response pattern is changed a small number of times in order to change the directivity in a trial, such as beam scanning.

V. NUMERICAL EVALUATION

We evaluate the characteristics of the EPEM algorithm and the EPEMD algorithm from the viewpoint of error reduction.

We designed the directivity pattern as shown in Figure 8. We then converted the directivity pattern into the detectable area using Equations (9) and (13). The detectable area is shown in Figure 9. In the simulations, we assume that the maximum value of the $|S_{sum}(\theta, t)|$ is 2 and that the maximum detectable range R' is 10 m. As mentioned in Section IV, it is necessary to change the directivity in a detection trial such as beam scanning in order to detect targets over a wide area. However, for the evaluation of the position estimation characteristics of the algorithms, only one fixed directivity pattern is simulated. The parameters of the transmitter are shown in Table III. Considering that targets exist in the near field, the width of the transmission array is set to 0.1 m. We simulate two cases. In Case 1, the target is located at (0,9) [m], which is a relatively long distance from the sensors. In Case 2, the target is located at (0,2) [m], which is short. For the evaluation of the estimation errors, we use the variance of the distribution of the estimated positions, which are used in Section III-B, as the performance measure. The other simulation parameters are as listed in Table I.

The results for the variance are shown in Table IV. These variances are derived from the distribution of the estimated positions. For example, the obtained distributions in Case 1 are shown in Figure 10,11. From Table IV, in the case of the EPEMD algorithm, the variance of both the x- and y-directions can be reduced compared to the EPEM algorithm. Moreover, in the case of a long distance, the reduction in variance is large compared to the case of a short distance. In particular, the variance in the x-direction, which is the direction of the arrangement of the receivers, can be decreased significantly.

VI. CONCLUSION

We considered the position estimation algorithm for targets that exist in the near wide area. A problem exists in that the position estimation error in the direction of the arrangement of receivers is large if the receivers are arranged along a straight line. In order to reduce this error, the proposed EPEMD algorithm uses the target existence probability, which is calculated based on range, and the directivity information of the transmitter. Computer simulations revealed that the proposed algorithm achieved low errors. Moreover, the error in the direction of the receivers arrangement was effectively reduced as intended.

As presented in Table IV, the variance in the y-direction is very small, which indicates that the multiple networked sensors have significant potential for application. Compared to the error in the y-direction, the error in the x-direction is relatively large. As the future work, we will continue to reduce this error. We will first find a suitable directivity pattern for the EPEMD algorithm. Next we will apply the

 ISBN:978-1-61208-113-7

Table IV
MEAN AND VARIANCE VALUES

	Target position [m]	Method	Var[$\hat{x}$]	Var[$\hat{y}$]
Case 1	(0,9)	Conventional	0.240	0.00237
		EPEMD	0.0339	0.00179
Case 2	(0,2)	Conventional	0.0139	0.00216
		EPEMD	0.0115	0.00227

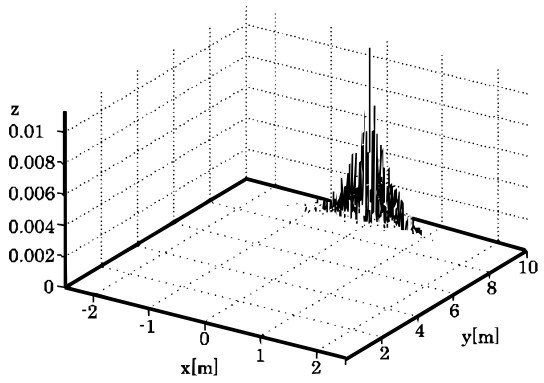


Figure 10. Distribution of estimated positions (Conventional, Case 1)

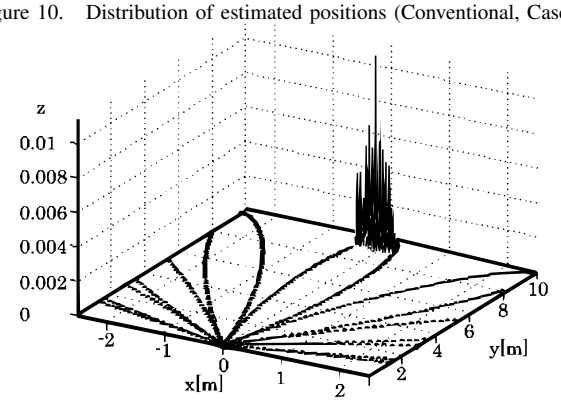


Figure 11. Distribution of estimated positions (EPEMD, Case 1)

reflected signals, which are often dealt with as multipath signals, to the EPEMD algorithm. The reason is that the multipath can surround the target whereas the receivers cannot surround the target.

ACKNOWLEDGMENTS

The present study was supported in part by a Grant-in-Aid for Young Scientists (B) and by the Telecommunications Advancement Foundation.

REFERENCES

[1] M. Klotz and H. Rohling, "A high range resolution radar system network for parking aid applications," *International Conference on Radar Systems*, May 1999.

[2] H. Rohling, A. Hoess, U. Luebbert, and M. Schiementz, "Multistatic radar principles for automotive radarnet applications," *German Radar symposium 2002*, Sep. 2002.

[3] A. Boukerche, H. Oliveira, E. Nakamura, and A. Loureiro, "Localization systems for wireless sensor networks," *Wireless Communications, IEEE*, vol. 14, no. 6, pp. 6 –12, dec. 2007.

[4] V. Ramadurai and M. L. Sichitiu, "Localization in wireless sensor networks: A probabilistic approach," *Proceedings of the International Conference on Wireless Networks, ICWN '03*, pp. 275–281, June 2003.

[5] D. Niculescu and B. Nath, "Ad hoc positioning system (aps) using aoa," *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, vol. 3, pp. 1734–1743, April 2003.

[6] N. B. Priyantha, A. K. Miu, H. Balakrishnan, and S. Teller, "The cricket compass for context-aware mobile applications," *Proceedings of the 7th annual international conference on Mobile computing and networking*, pp. 1–14, July 2001.

[7] S. Simic and S. S. Sastry, "Distributed localization in wireless ad hoc networks," no. UCB/ERL M02/26, 2002.

[8] Y. Chengyou, X. Shanjia, and W. Dongjin, "Location accuracy of multistatic radars (trn) based on ranging information," *Radar, 1996. Proceedings., CIE International Conference of*, pp. 34–38, oct. 1996.

[9] A. Hoess, H. Rohling, W. Hosp, R. Doerfler, and M. Brandt, "Multistatic 77GHz radar network for automotive applications," *ITS World congress 2003*, Nov. 2003.

[10] R. Mende, "A multifunctional automotive short range radar system," *German Radar Symposium 2000*, Oct. 2000.

[11] H. Hatano, T. Yamazato, H. Okada, and M. Katayama, "Target position estimation using MMSE for UWB IPCP receiver," *The 5th international conference on intelligent transportation systems telecommunication*, no. 45-3963283501, Jun. 2005.

[12] D. Oprisan and H. Rohling, "Tracking system for automotive radar networks," *RADAR 2002*, pp. 339–343, Oct. 2002.

[13] M. Klotz and H. Rohling, "24 GHz radar sensors for automotive applications," *International Conference on Microwaves, Radar and Wireless Communications*, vol. 1, pp. 359–362, Sep. 2000.

[14] H. Hatano, T. Yamazato, H. Okada, and M. Katayama, "A simple estimator of multiple target positions for automotive short range radar networks," *IEEE vehicular technology conference 2007-spring*, pp. 2511–2515, Apr. 2007.

[15] H. Hatano, T. Yamazato, and M. Katayama, "Automotive ultrasonic array emitter for short-range targets detection," *IEEE international symposium on wireless communication systems*, pp. 355–359, Sep. 2007.

[16] M. E. Russell, C. A. Drubin, A. S. Marinilli, W. G. Woodington, and M. J. D. Checcolo, "Integrated automotive sensors," *IEEE Trans. Microwave Theory and Techniques*, vol. 50, no. 3, pp. 674–677, Mar. 2002.

 ISBN:978-1-61208-113-7

A Dynamic Bandwidth Allocation Scheme for Interactive Multimedia Applications over Cellular Networks

Kirti Keshav
Samsung India Software Operations Pvt Ltd.
Block B, No 66/1 Bagmane Tech Park
C V Raman Nagar, Bangalore, India
kirtik@ece.iisc.ernet.in

Pallapa Venkataram
ECE Department
Indian Institute of Science
Bangalore, India
pallapa@ece.iisc.ernet.in

***Abstract*—Cellular networks played key role in enabling high level of bandwidth for users by employing traditional methods such as guaranteed QoS based on application category at radio access stratum level for various classes of QoSs. Also, the newer multimode phones (e.g., phones that support LTE (Long Term Evolution standard), UMTS, GSM, WIFI all at once) are capable to use multiple access methods simultaneously and can perform seamless handover among various supported technologies to remain connected. With various types of applications (including interactive ones) running on these devices, which in turn have different QoS requirements, this work discusses as how QoS (measured in terms of user level response time, delay, jitter and transmission rate) can be achieved for interactive applications using dynamic bandwidth allocation schemes over cellular networks. In this work, we propose a dynamic bandwidth allocation scheme for interactive multimedia applications with/without background load in the cellular networks. The system has been simulated for many application types running in parallel and it has been observed that if interactive applications are to be provided with decent response time, a periodic overhauling of policy at admission control has to be done by taking into account history, criticality of applications. The results demonstrate that interactive applications can be provided with good service if policy database at admission control is reviewed dynamically.**

Keywords-Cellular networks; interactive applications; dynamic bandwidth allocation.

I. INTRODUCTION

As the bandwidth available to the wireless network users increases, applications (APPs) are becoming more bandwidth hungry. Apart from the bandwidth, other main requirements for current day interactive applications are tight temporal relations, less loss rates and energy efficient data transfer for battery operated hand held devices. This is because session time of interactive applications tends to be longer with intermittent data transfer and in such cases mobile phone's energy consumption should be minimized. Various radio resource management (RRM) techniques has been developed for provisioning QoS for wireless cellular networks. Usually network operators chooses proper access methods and proper RRM policies for multi mode mobile phones.

Interactive traffic classes are meant for bursty, intermittent data transmission. During interactive session, key factor is the response time to users at both the ends. Amount of data transfer could be huge, but it would be bursty in nature. Examples of such applications include instant messenger for voice, chat & video, placeware (conference), netmeeting whiteboard, video messaging, online trading or transactional data client/server applications (such as SAP's, Peoplesoft's and Oracle's), and telnet. Another example of interactive applications is Navigation using off board method, in which the map to be presented on user's screen depends upon, which direction user would take on road. Apart from these applications, many time critical application such as in healthcare sector wherein the doctor needs to instruct interactively for surgery remotely to his colleagues or systems at far off distance. In summary, the main characteristic of interactive class traffic is the strict response time requirement with interactive class traffic while other traffic types are not so delay sensitive.

There are major complexities involved for running real time interactive multimedia applications over cellular networks. Few examples are given here. As the network delays are not predictable, many time critical applications like online trading can suffer and can cause financial losses to users. Similarly for interactive wireless games, if bandwidth and tight temporal requirements are not met, the whole gaming experience can be of very poor quality. For interactive news, if temporal relations between audio/video are not maintained, then it can lead to poor user experience.

Apart from the requirement of necessary required data rate, temporal requirements and necessity of mitigating effects of wireless problems such as fading, frequent handovers due to mobility pose the major challenge for seamless interactive multimedia experience. With the increasing demand for multimedia services in wireless networks, a great deal of attention needs to be paid to resource allocation for time critical interactive multimedia application data and on ways to provide seamless multimedia access in the next generation mobile communication networks. Problem of intelligent usage and allocation of the available bandwidth

 ISBN:978-1-61208-113-7

in a wireless environment is still a challenge due to client mobility and radio born errors.

In the remaining of this sections, QoS requirements of interactive applications and problems faced by interactive applications in cellular network environment are discussed. Following the discussion about related work in Section II, in Section III, the proposed bandwidth allocation scheme is pre- sented by giving details of components present in base sta-‘ tion. After providing detailed algorithm for selecting pri- orities according to host of factors to provide for reliable service to interactive applications, in Section IV, the simulation environment is presented. Thereafter, details of simulation procedure employed in designing of discrete event simulator are presented in Section V. Modelling aspects for various technologies such as LTE (Long Term Evolution), WCDMA (Wideband Code Division Multiple Access) are presented in Section VI. Sections VII and VIII contain details of results and conclusion, and finally, in Section IX, direction for future works is provided.

A. QoS requirements for interactive applications

Interactive class applications mainly comprise of a human or a machine or a remote server at both the ends of communication link and having a response time constrained data transfer. It is characterized by the request response pattern of the end user. Also interactive applications are generally foreground operations in which user waits for the operation to complete before proceeding further and have very tight response time (latency) requirements. Multimedia data can include text, graphics, images, audio and video in various combinations. By strict temporal relations it is meant that the different data types must be received and presented to user as per critical timing relationships. An entity at the destination is usually expecting a response message within a certain period of time. Therefore, the round trip propagation delay (RTD) and delay jitter are the key requirements for such applications [9]. The examples of typical interactive applications such as interactive games and interactive class room illustrates such requirements. For example, interactive games are the one which use the network to interact with other users or systems. Although requirements would depend upon specific application, bandwidth, delay, delay jitter, response time are the key parameters. Many interactive applications try to exchange high volumes of data, but demand very short delays and response time of 100-250 ms is a typical requirement.

B. Problems in using interactive applications over cellular networks

The data encoding methods affect the experience of interactive applications and proper encoding methods should be choosen carefully. Various advances has occurred in data representation schemes for multimedia (MPEG-2, MPEG-4 etc). MPEG-1 and MPEG-2 employ frame based coding techniques in which each rectangular video frame is treated as a unit for compression. The main concern was high compression ratio and satisfactory quality of video under such compression techniques. MPEG-2 had small interaction and is therefore not much helpful for interactive features of communication over cellular communication. MPEG-4, besides compression enables features required for user interactions. It adopts to a new object based coding approach - media objects are now entities for MPEG-4 coding. Media objects can be either natural or synthetic. The bandwidth requirement vary from 5 Kbps to 10 Mbps, so MPEG-4 is highly suitable for interactive multimedia.

To support various kinds of quality of service (QoS) requirements for interactive multimedia in wireless networks, resource provisioning is a major issue. Call admission control (CAC) and dynamic bandwidth management are such provisioning strategies to limit the number of call connections into the networks and to down grade other low priority classes in order to reduce the network congestion and call dropping.

In wireless networks, another dimension is added: call drop due to high users mobility. A good CAC scheme has to balance the call blocking and call dropping in order to provide the desired QoS requirements. Reservation of resources in the network is required to help smooth working of interactive multimedia. This also requires coordination amongst all parties such as admission control, policing, etc.

II. EXISTING RELATED WORKS

One can find lots of work in the literature and standards which talk about providing QoS to various classes of traffic in wireless networks.

One of the most popular strategies for wireless mobile multimedia networks which serve different types of customers with differing bandwidth requirements is the reserve channels (RC) CAC strategy. Call admission control schemes can be categorized based on their handoff-priority policy or queuing priority schemes [5] [6].

1. Guard channel (GC) schemes: In this type of scheme, some channels are reserved for handoff calls. Four different types of CAC schemes have appeared in the literature: Cutoff priority scheme, fractional GC scheme, rigid division-based scheme and new call bounding scheme.

2. Queuing priority (QP) schemes: In this type of scheme, calls are accepted whenever there are free channels. Depending on the approach, new calls are blocked and handoff calls are queued, vice versa, or all calls are queued and the queue is rearranged based on certain priorities.

The work presented in this paper falls conceptually within the second category, that is, of the QP schemes, as in our CAC scheme, both handoff calls and the new calls originating from within the cell are accepted if enough free channel bandwidth exists to accommodate them, and no portion of the bandwidth is restricted for access of either

 ISBN:978-1-61208-113-7

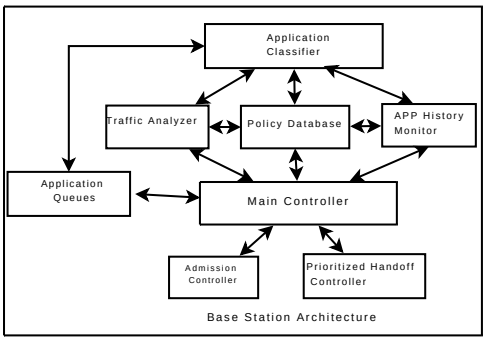


Figure 1. Architecture of bandwidth allocation scheme at base station

type of call (therefore, our scheme is conceptually similar to [8]).

III. Proposed bandwidth allocation schemes

In this paper, we propose a scheme to be considered at network side, to effectively allocate and maintain bandwidth in such a way that QoS requirements of various applications do not suffer and at the same time, deserving interactive applications get their bandwidth allotment easily and the network maintains the interactiveness with high fidelity.

A. Architecture of the bandwidth selection scheme at a base station

The architecture to be used by the base station (or Node B, e-NodeB, whichever is applicable based on communication system) to implement proposed bandwidth selection scheme is depicted in Figure 1 and as can be observed the main components are Application History Monitor, Traffic Analyzer, Application Classifier, Policy Database, Admission Controller, Prioritized Handoff Controller and Application Queues.

In the proposed system, base station maintains the QoS for interactive applications using admission control and traffic shaping operations based on dynamic policies which are derived from history of traffic using traffic analyzer and application history monitor component. Based on regular updating of policy database, intelligent maintenance of application queues, admission control and prioritized handoff control operations, an optimal traffic level is maintained in the system.

Here is the brief description of these components.

Application Classifier: Whenever a user invokes an application, application would request a particular QoS for channel access. Based on QoS parameters of these APPs, and other related parameters applications are classified at two levels - group category level and intra-group category level. Group category level is the broad classification while in intra group category, applications are categorized further inside groups and are assigned priorities.

Application Queues: This component stores the the various arriving applications and considers restrictions such has whether application QoS allows for waiting and channel capacity.

Traffic Analyzer: This component monitors the traffic over a long period of time and maintains the statistics such as application wise drop rate, degradation of QoS, etc. This component along with application history monitor, helps in predicting the application behaviors in better way and hence helps in adaptation of BSC according to current traffic profile.

Policy Database: This component has a database of dynamic policies, which helps in working of admission controller and handoff controller. Policies are dynamic in the sense that a more optimized set of policies are used dynamically based on result of monitoring of traffic by the traffic analyzer and application history due to previous set of policies over a period of time. This means that effect due to one set of policies is analyzed and based on feedback, the policies are updated dynamically for new arriving applications.

Main Controller: This is the central controller in the base station which maintains the central thread in the system. It negotiates with the core network component such as RNC, MSCs and other BSCs. It coordinates application queues, policy database, admission controller and handover controller.

Admission Controller: This component decides the fate of applications when they arrive. If channel doesn't have sufficient suitable amount of bandwidth, when application arrives, a decision to block or drop the call is taken here. This component works in tandem with policies component and main controller. Specially for interactive applications, more restricted admission control mechanism is used. Even if manageable bandwidth is available, an interactive application is admitted only after making sure the current application would survive the predicted scarcity of bandwidth in near future based on experience from previous application history.

Handover Controller: This component takes a decision regarding handover of current application to another cell, if current cell is totally occupied and neighbor cell can provide the required QoS. It also checks the direction of movement (to be negotiated through the main controller from other BSCs and RNC), serving cell signal strength, neighbor cell capabilities, application history for making the appropriate handover commands to users. This feature is very helpful especially in off-board Navigation, wherein application has to pre-download the map from server, based on the prediction of route which user might adopt.

Application History Monitor: This component monitors at detailed level, the history of application calls at the local level (BSC).

In this scheme priorities are assigned in a very dynamic way. Proposed scheme involves monitoring of various call performance parameters in every fixed interval and after

 ISBN:978-1-61208-113-7

Algorithm 1 Algorithm of proposed scheme

1: Begin
2: **Input: Application is to be dequeued from the APP wait List for admitting in system**
3: **Output: APP to admitted - Yes/No ?**
4: Initial priorities assigned to various APP categories based on APP characteristics
5: At every departure, priorities are reviewed
6: APP with maximum priority are chosen
7: **if** New APP doesn't meet bandwidth Requirement **then**
8: Do not drop the APP, wait for next departure. max wait time depend on APP.
9: **else**
10: Admit the APP
11: **end if**
12: Next admission decision taken at next departure
13: End

Algorithm 2 Simulation - Priority review algorithm

1: Begin
2: **Input: Periodic review timer has expired**
3: **Output: New updated policies**
4: Check and consult traffic analyzer, policy database, APP history monitor, interactive application response time.
5: Update priorities to be used in APP Wait Queue accordingly.
6: End

observing key characteristics a fresh decision regarding priorities for the next interval is taken. These priorities are then taken into account for decisions such as admission, dropping, reducing bandwidth of already accommodated application. Example of various decision factors which can be used in deciding priority are:

a) History of traffic pattern during the day: Traffic pattern in a given time interval of the day.

b) Geographical Area: (i) e.g., if it is hospital - voice traffic is generally given highest priority, however interactive health care applications are also taken care with more priority. (ii) for residential area, applications such as interactive games and shared video news are given more priority.

c) For office areas, application flows involving stocks, email, video conferencing are given high priority.

d) In the rural areas, interactive class room and other applications basted on history are decided to be of higher priority.

e) While in UMTS case, priorities are changed dynamically based on interference, in case of LTE, priorities of applications are updated dynamically based on location in the cell (cell edge users etc). For GSM cells, for various application types, differentiation is not done within cell.

Thus in proposed scheme, always a fare treatment to applications is done by network. Our scheme is shown, via an extensive simulation study comparison and a conceptual comparison, to clearly excel in terms of QoS provisioning to users for interactive applications while balancing overall network parameters such as overall reduction in cellular interference, overall throughput, edge-user throughput, cell user blocking probability and cell capacity . To the best of our knowledge, this is the first work in the relevant literature where such an approach has been proposed.

IV. Simulation Environment & Procedure

For simulation, discrete event simulation methodology has been employed. In the simulation, application arrivals are assumed to be random, and are considered as poisson arrivals. When an arrival occurs, it is classified into one of the application types as mentioned in the table 1 and whether it is same cell generated or its a handover-ed application from another neighboring cell is decided.

Application category is decided randomly using a uniform random variable. Different application types are assigned higher/lower rate according to the arrival rate desired for simulation. For example if APP1 type of application are desired to be of higher arrival rate type, then they are assigned more range in the uniform random variable output while determining their probability of arrival.

Algorithm 1 describes the procedure of admitting new application in the system. Based on application's type, characteristics such as bandwidth required and time it will stay in the cell are calculated. In the simulation, while bandwidth required are hard coded, stay time in cell is estimated using exponential distribution, with each application type having different mean stay time. During simulation, interactive applications are assigned longer call duration as shown in Table 1.

After determining application characteristics, available bandwidth in the channel is checked by the cell's bandwidth allocation module. If this module detects that it can accommodate the currently arrived call without affecting its current application performance, then the call is admitted. If however, bandwidth allocation module finds that the arriving call cannot be admitted, then the call is put into queue. As shown in Algorithm 1, applications are not immediately dropped if bandwidth is not sufficient at the time of new application arrival, rather a fresh assessment of situation is done at every departure of application to know if the leaving application has freed the necessary bandwidth in which new application can be accommodated. Each application category has separate queue having different sizes, with sizes defined by mean bandwidth requirement of respective application category. Each of this queue is assigned with a priority, which depends on many factors such as response time, application's history and criticality of application. Each of

 ISBN:978-1-61208-113-7

Table I
CHARACTERSTICS OF APPLICATION TYPES

	Reqd APP bandwidth(Kbps)	Mean call duration	Max wait time	Percentage of App	Example
APP1	20	5	1	40	Voice Call
APP2	40	2	2	10	InterAct Video News
APP3	10	1	1	20	InterAct Share Mkt Trading
APP4	60	10	1	30	InterAct Wireless Games

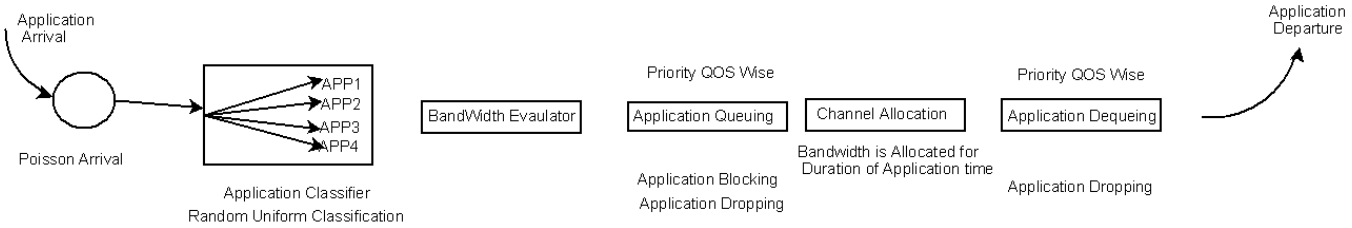


Figure 2. Simulation Model

the application type has a maximum wait period, on whose expiry, if application is still not scheduled to be taken into channel, the application is dropped.

Now whenever, an application leaves the channel, the bandwidth allocator sees if some applications are pending in the queues. It checks the queue priority wise, with higher priority given to more real time interactive application categories and with those APPs having less wait time. Hence by controlling priority of queues and wait time of different queue types, QoS for interactive applications can be guaranteed. Applications flow characteristics are depicted in Table 1. As described above, simulation environment is simulated by performing simulations as per the simulation model depicted in Figure 2. Special care is taken for interactive applications. In today's smart phone, various types of applications are simultaneously running. However it is the interactive applications, which demands most real time behavior from network. For this reason, as described in Algorithm 2, policies are revised periodically based on consultation from traffic analyzer, APP history monitor and interactive application specific response time. Accordingly whenever the network sees that new application type is of interactive type, the main controller in our proposed scheme, learns from previous experience by discussing with App history monitor to check if the available bandwidth with Network is sufficient for future needs of this type of Interactive application. So in essence, apart from assigning higher priorities to interactive applications, our scheme takes care of smooth running of interactive applications by learning in advance about the future need of similar application. This learning of behavior of various applications is done for all users of Network and a central database about application history is maintained. Every e-NodeB/RNC consults this database and takes action at NodeB level at the time of admission of new interactive applications.

V. MODELING ASPECTS

The actual radio resource management (RRM) algorithm definitely depends upon the technology been used at physical layer. For example, if a CDMA- or WCDMA-based (3G) system is considered, then, the fact that system capacity of radio channels (e.g., in 2G systems), but is rather affected by so called "Soft Limited" interference based CDMA behavior has to be kept as a criteria for RRM algorithms. Similarly, if OFDMA is considered at physical Layer (4G, LTE), then apart from usual parameters such as users requirements on the radio and the channel condition, the location of the user in the target cell becomes a significant factor. This is because in LTE, because of OFDMA, there is almost nil intra cell interference, but a high inter-cell interference exist as cell edge users share the same bandwidth in neighboring cell. So to provide good QoS to interactive applications in terms of upper limit on response time, application specific needs have to be taken care by bandwidth allocation and traffic shaping schemes in a cross layer way.

In WCDMA case, the parameters the energy per data bit to effective noise power density ratio $\frac{E_b}{N_t}$ and data rate R_b help in decoupling load control with physical layer parameters [3], [4]. It turns out that relationship between data rate and resource consumption is non linear. This is because resource consumption also depends on QoS requirements [4].

Uplink direction:

$$\left(\frac{E_b}{N_t}\right)_j = \frac{W/R_{bj}.\hat{P}_j}{(1-k).\sum_{i\neq j}\hat{P}_i + I_{other} + I_{th}} \qquad (1)$$

W - The chip rate of the system.In UMTS the rate is @=3.84 MChips/S

$\hat{P}_i, \hat{P}_j$- the mean received powers from users i, j respectively.

k - Describes the influence of interference reduction schemes, such as multi user detection (MUD)

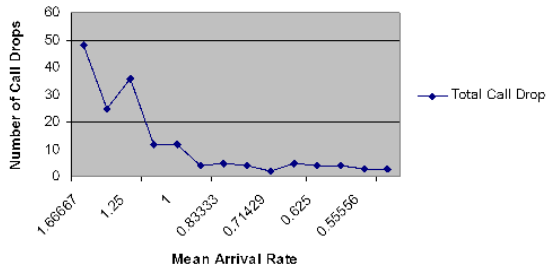


Figure 3. Call Drop Performance in this Scheme

$Iown - \sum_{i \neq j} \hat{P}i$ - the interference from other users in the same cell(cell of user j)

I_{other} - The interference from all users in other cells(other than user j^{th} cell) where the signal is not destined to the referenced Node B (Base station of user j).

I_{th} - The thermal background noise.

The current uplink load can be estimated using $\eta_{ul} = \frac{I_o - I_{th}}{I_o}$, where I_o is the total power received at Base Station.

Downlink Direction:

In the downlink direction, similar analogous results can be derived with the difference that in downlink, one Node B transmits data to many mobile users (mobile stations) and reference point for the downlink investigations is at the user j. The current downlink load can be estimated using $\eta_{dl} = \frac{I_o - P_p}{I_o}$, where I_o is the total power received at base station.

Similarly for LTE, the model followed is similar to "Softer Frequency Reuse based Resource Scheduling Algorithm", as mentioned in [7], in which cell edge users are having greater probability to use the frequency band with higher power and the cell center users are having the higher probability of using frequency band with lower power.

VI. Results and Discussion

Using the above mentioned method of periodic overhauling of policy for admission and traffic shaping for dynamic bandwidth management, after simulations it is seen that apart from being able to provide better QoS, response time to interactive applications, the over all call drop is also kept in check (Figure 3). Since the proposed scheme involves monitoring of various call/application performance parameters every fixed interval and since after observing key characteristics a fresh decision regarding priorities for the next interval is taken, the result obtained suggest that for interactive applications, dynamically adapting of policies is must in order to give good experience to interactive application users while maintaining the QoS of background services under check.

VII. Conclusion

As we have seen in above sections, the proposed bandwidth allocation scheme is an efficient one and when compared to other works, its performance is proved to be better. This shows that in order to provide time critical response to interactive users in reliable way, bandwidth management policies should be reviewed periodically and should take into account the response time, traffic pattern, previous call drop rates, application history (which applications are critical and which flows have given maximum profit in past to network operator).

VIII. Future Work

We are working on design of comprehensive framework to include all aspects of radio resource management such as admission control, scheduling, policing, etc., so as to provide best service for interactive applications.

References

[1] Y. Xu, H. Liu, and Q. Zeng, "Resource management and QoS control in multiple traffic wireless and mobile Internet systems", 14th International Conference on Computer Communications and Networks, vol. 17, no. 19, pp. 125-130, Oct. 2005.

[2] A. Kind, X. Dimitropoulos, S. Denazis, and B. Claise, "Advanced Network Monitoring Brings Life to the Awareness Plane" IEEE Communications Magazine, vol. 46, no. 10, pp. 140-146, October 2008.

[3] 3GPP Specifications, Radio Resource Control (RRC) Protocol specification, 25.331, version 3.21.0, Release 99

[4] U. Bernhard, E. Jugl, J. Mueckenheim, H. Pampel, and M. Soellner, "Intelligent Management of Radio Resources in UMTS Access Networks" Bell Labs Technical Journal, vol. 7, no. 3, pp. 109-126, 2003.

[5] Y. Fang and Y. Zhang, "Call Admission Control Schemes and Performance Analysis in Wireless Mobile Networks" IEEE Transactions on Vehicular Technology, vol. 51, no. 2, pp. 371-382, 2002.

[6] H. A. Mohamed, "Call Admission Control in Wireless Networks: A Comprehensive Survey" IEEE Communications Surveys & Tutorials, vol.7, no.1, pp. 49-68, 2005

[7] S. Hussain, "Dynamic Radio Resource Management in 3GPP LTE", http://www.bth.se/fou/cuppsats.nsf/all/c858bf5b4979a6b3c1257552004262f6$file/Sajid_Hussain_MSc_Thesis_Report.pdf, last accessed on 1-11-2010

[8] C. J. Haung, W. K. Laib, and Y. L. Yanb, "A self-adaptive bandwidth reservation scheme for sectored cellular communications" Information Sciences, vol. 166, no. 1-4, pp. 127-146, 2004.

[9] B. Zheng and M. Atiquzzaman, "System Design and Network Requirements for Interactive Multimedia" IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 1, pp. 145-153, 2005

 ISBN:978-1-61208-113-7

Distributed TDMA MAC Protocol with Source-Driven Combined Resource Allocation in Ad Hoc Networks

Myunghwan Seo, Hyungweon Cho
Wireless Comm. Group
Samsung Thales Co., LTD.
Yongin, 449-885, South Korea
{*myunghwan.seo, hyungweon.cho*}*@samsung.com*

Jongho Park, Jihyoung Ahn, Bumkwi Choi, and Tae-Jin Lee
School of Information and Communication Engineering
Sungkyunkwan University
Suwon, 440-746, South Korea
{*tamalove, ahnjh, tigerghost, tjlee*}*@ece.skku.ac.kr*

***Abstract*—MAC protocols for multimedia traffic transmission in mobile ad hoc networks have been studied steadily. It is still challenging to design an efficient protocol due to limited resource and multi-hop transmission property. CSMA-based MAC protocol is not suitable for multimedia traffic transmission due to hop-by-hop contention and collision. TDMA-based MAC protocol is appropriate to transmit time-sensitive traffic but slot allocation and sharing of slot information mechanism is essential. In this paper we propose a novel TDMA MAC protocol for mobile ad hoc networks, which includes distributed resource reservation and resource state sharing mechanism. In addition, we present an efficient collision resolution method. We show that our proposed TDMA MAC is suitable for delay-sensitive data transmission via simulations.**

***Keywords*-ad hoc networks, MAC protocol, TDMA, QoS.**

I. INTRODUCTION

In the next generation wireless communication systems, it is likely that there will be an increasing demand for fast deployment of independent mobile devices, e.g., establishing decentralized and dynamic communication links for emergency operations, industrial process monitoring, and military networks. A mobile ad hoc network (MANET) is an autonomous collection of mobile nodes connected by wireless links. Since a MANET is a self-configuring network of mobile nodes, it does not have/need any centralized coordinator such as base stations (BSs) or access points (APs). Typically, routing functionality is incorporated into each mobile node in MANET, so that mobile nodes can communicate with one another over wide range of distance. In MANET, it is required to setup multi-hop paths in a proactive or reactive manner, and the routing functionality relies on Medium Access Control (MAC) layer. So it has the same problem as that in conventional wireless networks, e.g., hidden/exposed terminal problem. Moreover, it might be challenging to meet end-to-end performance requirements due to limited radio resource, path set up/management overhead or mobility.

To control the access to the medium for MANET, contention-based MAC protocol, i.e., IEEE 802.11 Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) MAC protocol, has been typically considered [1], [2]. In this approach, every mobile node has a fair chance to transmit data by the Binary Exponential Backoff (BEB) algorithm to avoid collisions with mobile nodes in a network. When the network traffic becomes heavy, overall network throughput might be degraded due to frequent collisions. Moreover, end-to-end delay may increase when information is transmitted over multiple mobile nodes along a path.

Time Division Multiple Access (TDMA) MAC protocol is an alternative way to resolve shortcomings of contention-based MAC protocol [3]. In this approach, time resource is slotted and these time slots are assigned to mobile nodes in an orthogonal way, which results in contention-free medium access. A main challenge of TDMA MAC is how to assign/schedule specific time slots to mobile nodes. In MANET, since there is no centralized coordinator, time slots should be reserved in advance or assigned in a distributed manner. In addition, all mobile nodes have to be time-synchronized to utilize regular time slots. If delay-sensitive and real-time data are to be communicated over multi-hop links, MANET has to support quality of service (QoS).

There have been some research work to design distributed TDMA-based MAC protocol for ad hoc networks [4]–[9]. In [4], Unifying Slot Assignment Protocol Multiple Access (USAP-MA) is proposed. It provides broadcast and unicast transmission of datagram and high capacity or low latency streams, sparse and dense neighborhood optimization, and the ability to scale up to large networks. However, explicit method to resolve collision during slot reservation is not provided. The node activation multiple access protocol (NAMA) [5], assigns time slots to mobile nodes based on their priorities. Although NAMA allows collision-free broadcast transmissions, it may not be able to control the number of

This work was supported by a grant-in-aid of SamsungThales, and by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010–(C1090-1011-0005)), and by Future-based Technology Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (20100020729).

 ISBN:978-1-61208-113-7

time slot allocations. In [9], Max Spatial Reuse Scheduling Algorithm (MARSA) is proposed. MARSA allows 1-hop distance neighbor nodes to communicate with other nodes in the same slot using directional antenna. In [8], cross layer design is provided for video streaming over wireless ad hoc networks using multichannels. The authors define Maximum Latency Rate (MLR) to provide differentiated traffic for TDMA structure. And congestion-aware routing protocol with congestion-aware metrics (MAC utilization and queue length of MAC) is proposed. GMAC [6] exploits the geographic positions of neighbors to construct a TDMA schedule. Although no negotiation phase is needed to assign time slots to nodes, GMAC relies on accurate location awareness, e.g., devices equipped with position awareness via a system such as Global Positioning System (GPS). In TDMA based multi-channel MAC (TMMAC) [7], every node negotiates which channel and time slot to use for data communications using IEEE 802.11 DCF, which results in frequent collisions when traffic becomes heavy.

In this paper, we propose a novel TDMA MAC protocol for MANET, which supports QoS and assigns time slots to mobile nodes in a distributed way. In the proposed MAC, each mobile node exchanges its routing information and resource allocation information to neighbors periodically in a pre-assigned time slot so that they can transmit network information without collisions. Using the resource allocation information from neighbors, each mobile node can reserve time slots for data transmissions without additional contention. In case of a reservation slot collision, our proposed MAC resolves the collision quickly by the preemption value without exchanging additional information. When a collision occurs, each node evaluates its own preemption value using neighbor information, and the node with higher preemption value occupies the collided time slot and the node with lower value tries reservation in another slot. The proposed MAC assigns priority to each node or traffic flow so that the node or the traffic flow with higher priority can reserve time slots before those with lower priority to support QoS.

The organization of the paper is as follows. The proposed MAC protocol including collision resolution mechanism is described in Section II. We simulate our proposed MAC in Section III. Finally, we make a conclusion in Section IV.

II. PROPOSED TDMA MAC

A. TDMA MAC Frame Structure

Our proposed TDMA MAC frame is composed of Network Information Broadcast (NIB), Resource Reservation (RR), and user data slots (see Fig. 1). We assume that the TDMA frame synchronization is provided by GPS, and that an NIB slot and an RR slot are pre-assigned to a specific node. So the number of NIB slot and an RR in a cycle are same as the number of nodes and NIB slot and an RR slot allocation relate to a node ID. These assumption is suitable for military, emergency, and private ad hoc network which

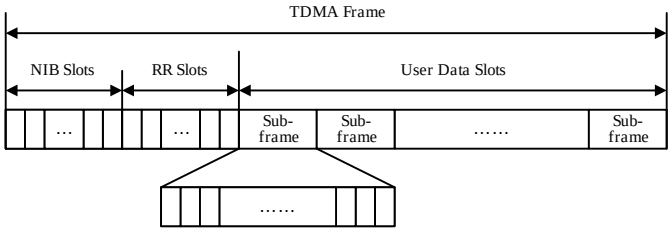


Figure 1. Proposed TDMA frame structure.

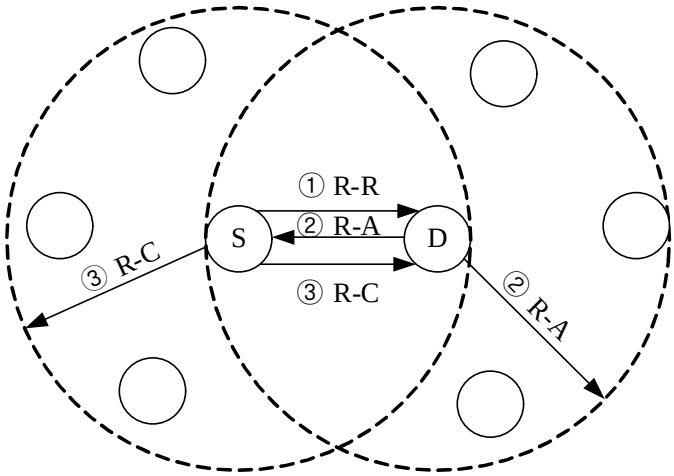


Figure 2. Resource reservation procedure during an R-R slot.

do not require self-organizing function. The NIB slots are used for broadcasting of routing information and resource allocation information. The broadcast message in an NIB slot is utilized to maintain the information of an ad hoc network. The routing information message can be different depending on an ad hoc routing protocol. If a reactive ad hoc routing protocol is used, the routing information may be route request or route reply. If a proactive ad hoc routing protocol is used, the routing information may be a hello message. In the NIB slots, resource allocation information is also broadcast too. The resource allocation information denotes the occupation status of user data slots. Each node in a network maintains the Resource Allocation Table (RAT). A node creates a resource allocation information message by referring to its own RAT. When neighbor nodes receive a resource allocation information message from a node it updates its own RAT. The RAT includes the resource allocation information for 3-hop neighbors to maintain user data slot occupation information of neighbors. The nodes over 3 hops do not interfere with one another, so they can use the same user data slot simultaneously.

In the RR slots, reservation messages are exchanged. The reservation is performed in a 3-way hand shake way. The source node who wants to send data transmits a Resource Request (R-R) message in the pre-defined RR slot. The R-

R message includes source ID, destination ID, sub-frame index, number of slots to use, and duration of slots to use in a sub-frame. A source node tries to reserve as many slots as the hop count of a path for source-driven combined multi-hop reservation. So hop-by-hop user data slot reservation is not necessary, which results in low end-to-end delay. When a node receives an R-R message, the node checks its own RAT and decides the requested user data slots are available. If the requested slots are available, the node transmits a Resource ACK (R-A) message including the information of the original R-R message. If they are not available, the node does not transmit anything. The source node waits for the R-A message before the R-A timer expires. If the source node receives the R-A message, the source node transmits a Resource Confirm (R-C) message including the information of the original R-R message. If the source node does not receive the R-A message before the R-A timer expires, the source node tries reservation again in the next TDMA frame. The nodes which overhear the R-A and/or R-C messages update their own RATs. The resource reservation procedure is depicted in Fig. 2.

A user data slot is composed of sub-frames. A user data slot is identified by a sub-frame index and a slot index in a R-R message. The slot index may be indicated by a bitmap of slots. So the sub-frame structure reduces the overhead of resource allocation information messages and resource reservation procedure. A node can request user data slots only in a sub-frame to prevent monopolization of user data slots. If a node wants to transmit best effort traffic, it requests slots in a sub-frame first, and if more slots are needed, the node requests slots in another sub-frame at the next TDMA frame. As a result, allocated slots for best effort traffic are slowly increasing.

The cycles for the transmission of NIB and RR messages can be defined appropriately. For example, if there are 8 nodes and the NIB cycle is 4 frames, node 1 and 2 use the NIB slots in the ith frame, and node 3 and 4 use the NIB slots in the $(i+1)$th frame. The cycles of NIB and RR messages have to be defined carefully by considering network characteristics.

B. Collision Resolution

Our proposed TDMA MAC protocol provides collision-free access unless 2-hop neighbors use the same time slot (user data slot). In the proposed TDMA MAC, reservation is made based on RATs containing resource allocation information from 3-hop neighbors to avoid duplicate use of the same time slot. However, RATs cannot immediately reflect dynamically changing topologies. So collisions may happen, and network throughput might be degraded due to simultaneous transmission in the same time slot. For instance, at the initial stage of network, nodes have little knowledge about their neighbors, and the time slots which seem to be unoccupied may already be in use. In order to resolve such collision,

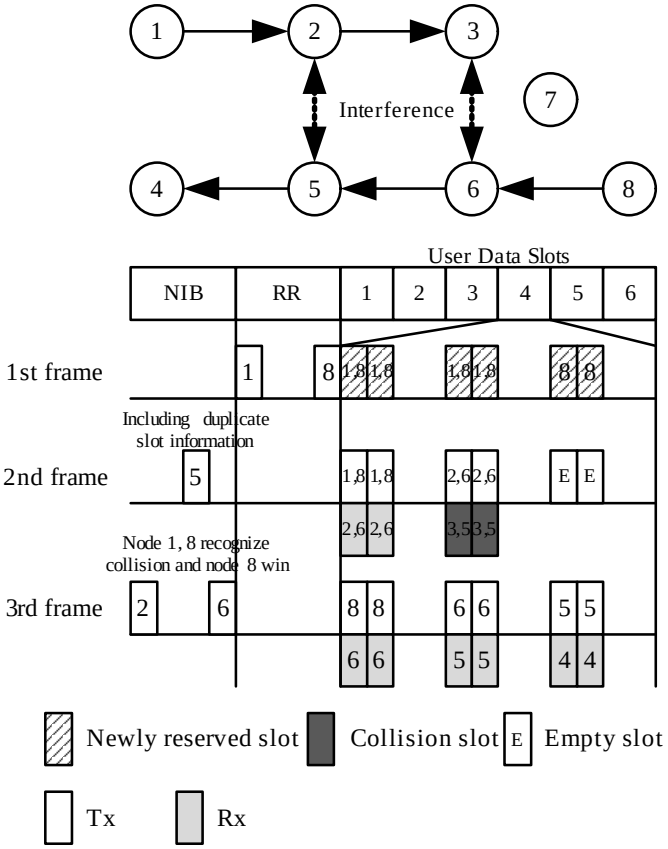


Figure 3. An example of collision resolution.

we present a simple and distributed method with minimum protocol overhead. It does not require any additional control message or negotiation between nodes. It refers to RATs exchanged in the NIB duration, which contains time slot schedule for upcoming user data slots.

A preemption value (V_{pre}) represents a priority for a certain time slot and a given node. A node with the highest V_{pre} value is allowed to use the collided time slot, and everyone else give up and look for another time slot which seems to be idle. So a node with the most important traffic is given a higher transmission priority. Providing fair transmission opportunity to the nodes with the same priority is also important. Thus we define V_{pre} as follows.

$V_{pre} = priority \& class \& \{(ID + T_{collision}) \% n\} \& ID$

where $\&$ and $\%$ denote bit level concatenation and modulo operation, respectively. $priority$ is the traffic priority, $class$ is the user priority, and ID is a uniquely identifiable address of a node. $T_{collision}$ is the time offset in microseconds from the beginning of a TDMA frame to the current time slot, and n is the number of nodes in the entire network. $T_{collision}$ can prevent a node or traffic from having relatively higher V_{pre} than any other node all the time, compensating unfairness in the transmission opportunity.

 ISBN:978-1-61208-113-7

Tx ID	Sub-frame index	Slot index bitmap	Duration
1	4	110011000000	20
8	4	110011001100	15

Figure 4. An example of RAT at node 5.

On receiving RATs, each node in a network checks for duplicate reservations. If duplicate reservations are found, G_{col}, a group of nodes which are involved in a collision, can be defined. Each node in G_{col} calculates the preemption values V_{pre} of every nodes in G_{col}. Only one node with the highest V_{pre} in G_{col} is eligible to use the collided time slot, and other nodes are supposed to make reservation on other time slots in the next frame.

Fig. 3 shows how collision resolution works. Assume that there are voice traffic demands from node 1 to node 3 and from node 8 to node 4, respectively. Since it is voice traffic, a source node manages reservation not only for one-hop neighbor but also for all the nodes in the route. Node 1 makes a reservation on the slots at the 4th sub-frame. On successful reservation, neighbors of node 1 (node 4 and 5 in this case) can overhear R-A or R-C message and update their RATs. At the same time, node 8 makes another reservation on the slots at the 4th sub-frame, and unfortunately both reservations use the same slots in the same sub-frame. Nevertheless, reservation request from node 8 is acknowledged because messages from node 1 and 2 are not heard to node 6. However data transmission from node 2 to node 3 interferes with data transmission from node 6 to node 5 (collision). It is not possible for node 5 and node 2 to realize collision immediately, because RATs for both traffic has not been distributed yet. In the NIB duration of the next frame, RATs are flooded one-hop further, and while updating RAT at node 5, node 5 can find the source of collision (An example of RAT records for node 5 is given in Fig. 4). In the third NIB duration, node 2 and node 6 transmit RATs to node 1 and 8, respectively. Then node 1 and node 8 finally recognize collision.

Now node 1 calculates V_{pre}s of node 1 and node 8, and so does node 8. By comparing the values they can decide which node wins the collided time slot. For example, assume that $priority = 3$, $class = 3$, $n = 8$, ID of Node 1 is 8453, and that of node 8 is 1584. And assume that $T_{collision}$ is 100μs. Then, V_{pre} is calculated as follows.
V_{pre}(Node 1)=3&3&$\{(8453 + 100)\%8)\}$&$8453 = 3318453$
V_{pre}(Node 8)=3&3&$\{(1584 + 100)\%8)\}$&$1584 = 3341584$
Since V_{pre}(Node 1) $\leq V_{pre}$(Node 8), node 8 wins the slot and node 1 should find idle slots at the next frame.

This collision resolution method can also be used to apply priority. If there is not enough slots to transmit a traffic flow, the node who wants to transmit checks its own RAT and tries to reserve the slots which are already used by a node or a traffic with lower priority (e.g., best effort traffic). Then the node with the already used slot recognizes that the slot collides and computes V_{pre}. The node will give up the slots due to lower priority.

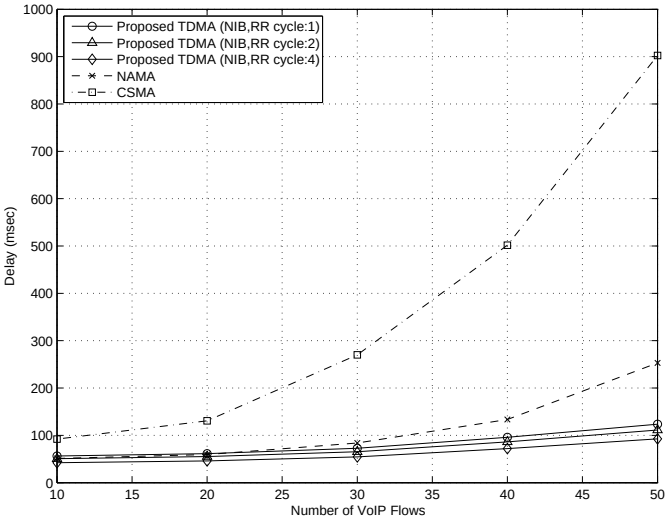


Figure 5. Delay of proposed TDMA, NAMA and CSMA.

III. Performance Evaluation

In this section, we present simulation results of the proposed TDMA MAC protocol, the existing NAMA [5] and the CSMA protocol. 100 nodes are distributed uniformly in 500mx500m space. The transmission range of a node is 200m. The PHY transmission rate is assumed to be 2Mbps. The Optimized Link State Routing (OLSR) is used for ad hoc routing protocol. The CSMA parameters are the same as those of IEEE 802.11b. The frame size of NAMA is 250ms and the data slot size is 0.4ms. The length of contention period to construct a contender set is 50ms. So there are 500 data slots in a frame. The frame size of the proposed TDMA MAC is 250ms when NIB and RR cycle is 1. The size of user data slots is 0.25ms. And the number of user data slots is 800. We consider Voice over IP (VoIP) traffic as delay-sensitive traffic. The rate of a VoIP flow is 6.4Kbps and the interval of VoIP packets is 30ms. And we use one FTP traffic for as background traffic. We vary the number of VoIP traffics from 10 to 50 and the NIB and RR cycle is 1, 2 or 4. The frame size of proposed TDMA MAC is changed with change of the NIB and RR cycle.

Fig. 5 shows the delay of VoIP traffic. The delay of the proposed TDMA MAC is slightly increased with the growth of VoIP traffics. The delay of the proposed TDMA MAC is below 150ms. As NIB and RR cycle becomes larger, the number of NIB and RR transmissions becomes smaller. So the delay of VoIP traffic improves. When defining NIB and RR cycle, mobility and traffic characteristics must be considered. The routing information and resource allocation

 ISBN:978-1-61208-113-7

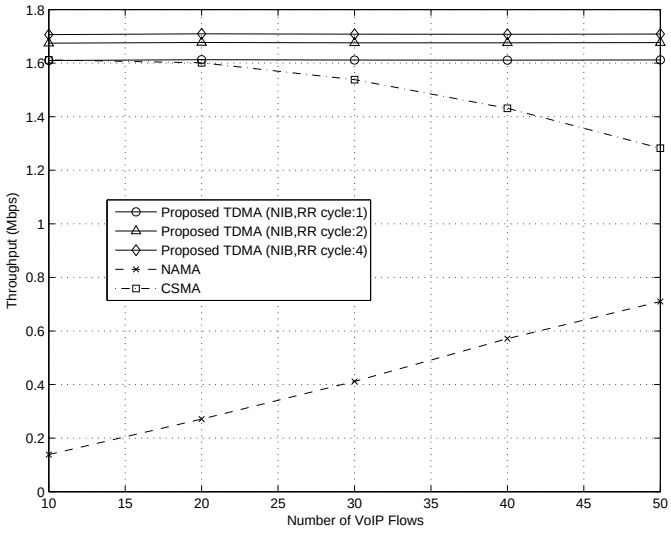


Figure 6. Total throughput of proposed TDMA, NAMA and CSMA.

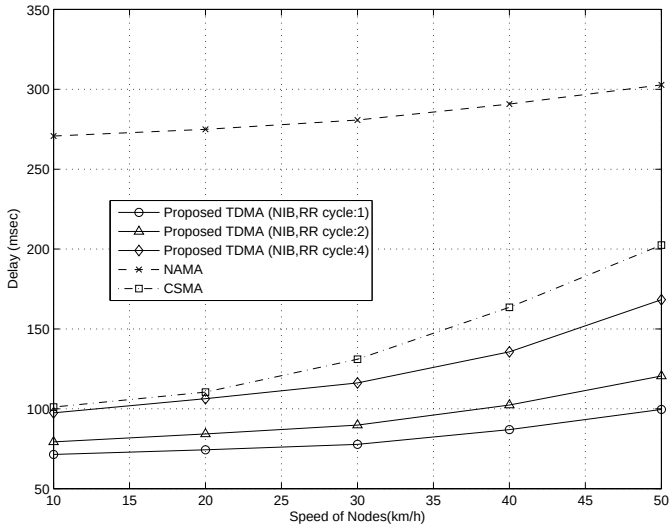


Figure 7. Delay of proposed TDMA, NAMA and CSMA with Mobility.

information in NIB is needed more frequently in the high mobility condition, and RR depends on traffic arrivals. The delay of NAMA is similar to the proposed TDMA protocol in low traffic load. However the delay of NAMA becomes larger than that of proposed TDMA as the number of VoIP traffics increases. The NAMA protocol may not schedule slots in parallel with traffic load, so the intermediate nodes in a routing path act as bottleneck nodes. The delay of CSMA increased rapidly with the growth of VoIP traffics. The CSMA can only accept less than 30 VoIP traffics if delay requirement is 300ms.

Fig. 6 shows total throughput of VoIP traffics and ftp traffic. The throughput of the proposed TDMA MAC is almost constant, which achieves the maximum throughput. As NIB and RR cycle increases, the throughput of the proposed TDMA MAC increases due to the relative growth of user data slots. The throughput of NAMA is relatively lower than other protocols. The NAMA protocol may not be able to differentiate traffic characteristics. So it allocates many slots to ftp traffic as well as VoIP traffics. The throughput of CSMA decreases rapidly as VoIP traffic increases. When there are many VoIP traffics, collision probability becomes high in CSMA.

Fig. 7 shows the delay of VoIP traffic with varying speed of nodes. We use random way point mobility model. The speed of nodes varies from 10 to 50Km/h. If a node reaches its destination, it waits a random time (0-30sec) and moves again. As NIB and RR cycle becomes larger, the routing information and slot allocation information are broadcast with long interval in the proposed TDMA protocol. The delay increase of the proposed TDMA protocol inherits from route change and collisions due to mobility. The delay of NAMA also increases as mobility increases. The main factor of delay increase is that the contender set becomes incorrect due to mobility. If the contender set is not perfect, there may be collisions. The delay of CSMA is influenced by routing protocol. Increasing delay of CSMA is caused by route failure.

IV. Conclusion

In this paper, we have proposed a new TDMA MAC protocol for MANET. In the proposed TDMA MAC, each mobile node broadcasts its routing information and resource allocation information to neighbors periodically in a pre-assigned time slot so that it can transmit network information without collision. Using the resource allocation information from neighbors, each mobile node can reserve time slots for data transmission without collision. In the proposed slot reservation procedure, three way hand shake mechanism lessens the hidden node problem. Although there might be reservation collision because of over 3-hop nodes, our proposed MAC resolves the collision by the preemption value without exchanging additional information. The simulation results show that our proposed TDMA MAC protocol is suitable for delay-sensitive data transmission in ad hoc networks.

References

[1] S.-T. Sheu and T.-F. Sheu, “DBASE: A Distributed Bandwidth Allocation/Sharing/Extension Protocol for Multimedia over IEEE 802.11 Ad Hoc Wireless LAN,” *in Proc. of IEEE INFOCOM*, pp. 1558-1567, Apr. 2001.

[2] Z. Yao, P. Fan, and Z. Cao, “An Enhanced CSMA-CA Mechanism for Multihop Ad Hoc Networks,” *in Proc. IEEE APCC/MDMC*, pp. 966-970, Aug. 2004.

[3] C. D. Young, “The Mobile Data Link (MDL) of the Joint Tactical Radio System Wideband Networking Waveform,” *in Proc. of IEEE MILCOM*, pp. 1-6, Oct. 2006.

 ISBN:978-1-61208-113-7

[4] C. D. Young, "USAP Multiple Access: Dynamic Resource Allocation for Mobile Multihop Multichannel Wireless Networking," *in Proc. of IEEE MILCOM*, pp. 271-275, Oct. 1999.

[5] L. Bao and J. J. Garcia-Luna-Aceves, "Distributed Dynamic Channel Access Scheduling For Ad Hoc Networks," *Journal of Parallel and Distributed Computing*, vol. 63, no. 1, pp. 3-14, 2003.

[6] J. Lessmann, "GMAC: A Position-Based Energy-Efficient QoS TDMA MAC for Ad Hoc Networks," *in Proc. of IEEE ICON 2007*, pp.449-454, 2007.

[7] J. Zhang, G. Zhou, C. Huang, S. Son, and J. A. Stankovic, "TMMAC: An Energy Efficient Multi-Channel MAC Protocol for Ad Hoc Networks," *in Proc. of IEEE ICC*, pp. 3554-3561, Aug. 2007.

[8] B. J. Oh and C. W. Chen, "A Cross-Layer Approach to Multichannel MAC Protocol Design for Video Streaming Over Wireless Ad Hoc Networks," *IEEE Trans. on Multimedia*, vol. 11, no. 6, pp. 1052-1061, Oct. 2009.

[9] Z. Guo and Y. Chen, "An Optimal Scheduling Algorithm in Spatial TDMA Mobile Ad Hoc Network," *in Proc. of MIKON*, pp. 1-5, Jun. 2010.

 ISBN:978-1-61208-113-7

Loss Differentiation and Recovery in TCP over Wireless Wide-Area Networks

Detlef Bosau
detlef.bosau@web.de

Herwig Unger, Lada-On Lertsuwanakul
Fernuni Hagen
{herwig.unger,lada-on.lertsuwanakul}@fernuni-hagen.de

Dominik Kaspar
Simula Research Laboratory, Norway
kaspar@simula.no

***Abstract*—The increasing speed and coverage of wireless wide-area networks (WWAN) has made technologies such as GPRS, UMTS, or HSDPA a popular way to access the Internet for both mobile and stationary users. However, depending on the scenario, WWAN can suffer from severe IP packet loss due to corruption, which is mistaken by TCP as an indication of path congestion. This paper presents an approach to solve the *loss differentiation problem* for the widespread scenario of wireless networks being used as access networks to the Internet. Our solution allows TCP to distinguish between congestion and corruption loss and to properly react to both phenomena. Loss differentiation is achieved by placing an assisting agent on the WWAN's base station, which replies a TCP sender in the wired Internet with a new type of acknowledgement. Without harming TCP's end-to-end semantics, these acknowledgements provide feedback about congestion on the wired path and corruption on the wireless path and support the sender in taking remedial action. Results from simulations indicate that our proposed corruption recovery algorithm significantly improves the TCP goodput. In addition, excessive RTO growth and pauses in the TCP flow that result from repeated packet corruption are considerably reduced.**

I. Introduction

Accessing the Internet and Internet-based services over wireless wide-area networks (WWAN) such as GPRS, UMTS, and HSDPA has become part of our everyday life. A typical scenario is given in Figure 1, which illustrates a mobile host (MH) attached to a wireless access network and communicating with a fixed host (FH) in the Internet. The wired Internet and the wireless access network are interconnected by a base station (BS).

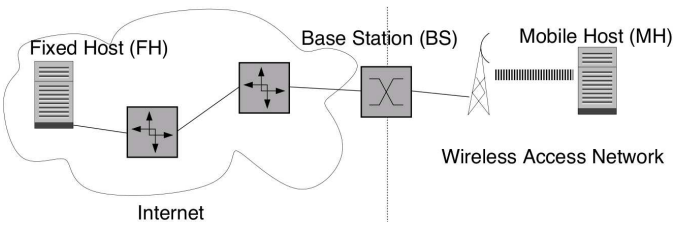


Fig. 1. Internet and wireless access network.

Due to the lossy nature of wireless links, TCP connections between FH and MH are faced with several difficulties.

- The *loss differentiation problem*: TCP implicitly assumes corruption-free links along the entire path, therefore packet corruption is mistaken as an indication of path congestion. As a consequence, TCP may not be able to fully utilize its fair share of the available path capacity.
- The proper *retransmission timeout (RTO) estimation* is known to be difficult [1][2]. The main challenge is caused by TCP's assumption of quasi-stationary round-trip times (RTT) [3][4]. A too large RTO causes inefficient usage of the available capacity, while a too small RTO may cause unnecessary retransmissions and congestion handling. The latter problem is also known as the "spurious timeout problem". Existing algorithms, such as TCP Eifel [5], detect spurious timeouts to correct an erroneously reduced congestion window, but they do not resolve corruption loss.
- TCP employs a Go-Back-N strategy for loss recovery, which is unsuitable when a packet requires more than one transmission attempt for successful delivery. TCP assumes corruption-free links, thus a packet retransmission is expected to be successful when a (perhaps erroneously!) recognized congestion is overcome. Further transmission attempts lead to increasing timeouts and unwanted pauses (refer to Section II-B for more details).

Particularly, the assumption of corruption-free links and quasi-stationary RTT is severely violated in WWANs. Depending on the actual scenario, packet corruption rates in WWANs can be arbitrarily close to 1. In addition, the service time for a packet on a WWAN link, and therefore the average RTT, may vary for several reasons, including varying cell load in cellular networks and therefore varying MAC delays, and varying channel or line coding if the network technology is able to adapt to different signal-to-noise ratios. As a motivating example, Figure 2 illustrates extreme RTT values (431 seconds at the maximum!) measured over HSDPA on a moving train.

In this paper, we decouple congestion control and pacing from reliable delivery by introducing an *assisting agent* on the base station, the "CLACK agent". This agent addresses the problems of loss differentiation and proper RTO estimation by providing the FH with a new type of acknowledgment. Hence, proper loss differentiation and clocking is achieved without harming the TCP connection's end-to-end semantics by ACK spoofing, connection splitting or introducing hard state into the network. The proposed CLACK agent can be considered as a "flow middlebox" in the sense of [6].

In addition, we propose a loss recovery strategy suitable for lossy networks, in which any packet corruption is signaled to

 ISBN:978-1-61208-113-7

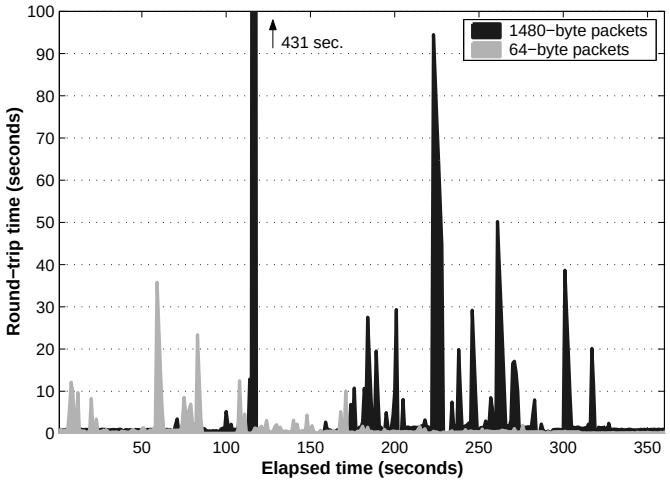


Fig. 2. Extreme round-trip times over an HSDPA access network during a train trip through the Norwegian countryside. The two experiments of small and large packet sizes were conducted at different times and locations.

the communication endpoints as soon as possible. Therefore, our mechanism provides a fast end-to-end retransmission of corrupted packets *before* the expiration of the RTO timer. In other words, our mechanism complements the fast retransmit of packets lost due to congestion by a fast retransmit of packets lost due to corruption.

The remainder of this paper continues with a discussion of related work in Section II. In Section III, a description of our CLACK approach follows. Section IV describes the details of our system model and our simulation environment. Simulation results are presented in Section V. Finally, Section VI concludes the paper and outlines our future research activities.

II. Related Work

The last two decades have seen a vast amount of related work in the area of wireless access to the Internet. Thus, only a selection of relevant papers can be addressed here for space limitations. For a more complete overview, the reader is referred to the work by Sonia Fahmy [7].

A. Local Recovery

In WWANs, the corruption probability for a given packet can be arbitrarily close to 1 and does not primarily depend on the technology in use but on the actual scenario. Packet corruption cannot be completely overcome by local recovery approaches, such as Snoop [8][9][10], which uses buffering and retransmission of lost packets at BS. Local recovery can increase the probability for a packet to be correctly delivered at the expense of increased delivery times. However, analogous to the "Two Army Problem", it is impossible to *reliably* deliver a packet over a lossy channel in *limited time*, or in a limited number of sending attempts.

A fundamental problem with all kinds of local recovery approaches and performance enhancing proxies (PEP)[11] is the eventual deletion of transient data from buffers and queues. Particularly in case of link outages and high corruption rates, packets buffered in some local recovery mechanism or PEP may stay forever and lead to unlimited head-of-line blocking, if they are not forcefully deleted (e.g., by a reasonably limited persistence). In this paper, we take the position that packet corruption should be handled by the connection endpoints and only limited effort should be spent on local recovery [12].

B. End–to–End Recovery

A second problem of lossy channels in TCP is the reliable end-to-end delivery of data when a lost packet cannot be recovered by a single retransmission. If a corrupted packet needs more than one transmission attempt for successful delivery, the individual retransmissions are started by retransmission timeouts, even if packet loss detection using duplicate and selective ACKs is employed. Because TCP employs cumulative ACKs and a Go–Back–N strategy for loss recovery, some of the yet unacknowledged data may have been successfully received and hence causes DupACKs during the retransmission process. Thus, DupACKs are not suitable for loss detection during the retransmission process. Refer to [13], Section 3.2 ("Fast Retransmit and Recovery") for details. Due to the retransmission timeout (RTO) backoff, which is used to rapidly increase the RTO to a sufficiently large value in case of a too small or not yet existing RTT estimate [14] repeated timeouts may lead to unwanted and annoying RTO growth and therefore pauses in the TCP transmission process.

To our knowledge, the problem of end–to–end loss recovery has received only little attention in the past. Approaches published so far consider loss recovery in Delay Tolerant Networks, e.g., [15] which reduce the number of retransmissions by adding redundant information to the data stream but do not primarily target recovery of actual packet loss.

C. Loss Differentiation

For the loss differentiation problem, we can identify three main approaches.

1) Some methods try to identify the reason for a packet loss by statistical means, e.g., [16]. The basic idea is to monitor a path's RTT and to infer from individual RTT observations, whether a "short time RTT average" exceeds a "long time RTT average", which is taken as an indication of congestion. Therefore, packet loss is recognized as congestion loss, while in the opposite case, packet loss is seen as corruption. Besides the end-to-end RTT, there are other statistics proposed for the same purpose of loss differentiation, e.g., interarrival times, interarrival time jitter etc.
   All of these approaches attempt to do an "ex post reasoning" in order to identify the actual reason for an individual observation, which may be caused by different reasons. For example, an increasing RTT can be due to a noisy channel and increasing *recovery delays*, or to a crowded cell with increasing *MAC delays*. This kind of ex post reasoning is never a proper application of mathematical statistics, particularly it does not yield the *true* reason for an *individual* packet loss.

 ISBN:978-1-61208-113-7

2) The *CETEN* approach by Eddy, Ostermann and Allman [17] does not focus on the individual packet loss but attempts to modify TCP's *Additive Increase Multiplicative Decrease* mechanism in order to accommodate the packet corruption rate along the path. As long as the packet corruption rate for the individual links is known, TCP is modified so that on average, the congestion window is halved when one congestion drop is observed. However, this approach is not feasible in a WWAN, where the actual packet corruption rate is unknown.[1]
3) There exist several flavors of rate-controlled TCP, which try to estimate the correct rate for a TCP flow and to accommodate accordingly. An example is TCP Westwood [18].
   The basic problem with rate-controlled approaches in wireless networks is that the service time for a Transport Block (TB) is neither known in advance nor does a sufficiently stable estimate exist. Therefore, these approaches are not suitable for WWAN.

III. The CLACK Agent / Approach

The basic idea of our approach is to place an assisting agent (the CLACK Agent) on BS in order to provide FH with proper pacing using Clock Acknowledgements (CLACKs), in addition to unaltered ACKs from the mobile endpoint MH. Figure 3 illustrates the components and nodes used in our approach. These components and their functionality are discussed in the following subsections.

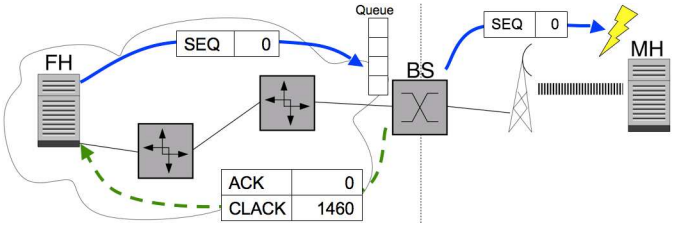


Fig. 3. For each TCP segment that left the wired path towards MH, the BS issues a CLACK (Clock Acknowledgement) back to FH. This example shows a *delivery failure* on the wireless link. Hence, the segment with sequence number 1460 is CLACKed together with the correct ACK number.

A. The Components on FH, BS, MH

1) The Receiver on MH: The receiver socket on the mobile host runs an ordinary TCP Reno without modifications.

2) The CLACK Agent on BS: The assisting agent on BS reacts to each TCP packet from FH in a similar way as an ordinary TCP receiver. However, acknowledgements issued by BS to FH consist of two values: the *CLACK* number, which indicates the next expected sequence number seen by the CLACK Agent *and* the *ACK* number, which is the next sequence number expected by FH. Please note, that TCP packets sent by MH are not simply forwarded by BS, but the acknowledgement information is conveyed to FH in the ACK field in packets sent by BS. Hence, proper pacing is achieved solely by the interaction of FH and BS. It must be pointed out that only unidirectional communication is discussed in this paper, bidirectional communication is left to future work, refer to Section VI.

Every packet received from BS is enqueued at BS or dropped in case of queue overflow. Any packet taken from the queue is forwarded to MH and acknowledged *regardless* whether the delivery attempt to MH is successful or not. This way, the queue from BS to MH appears as a part of the "Internet part" of the path. Queue overflow is handled on FH by the well known TCP congestion control [4].

For proper pacing, acknowledgements to be sent by BS are postponed until the acknowledged segment has left the outgoing queue of BS. While being busy with the delivery attempt, the outgoing interface of BS cannot accept new work. When the delivery attempt finishes, be it successful or not, the packet is released from the wired path (between FH and BS) and a new packet may be injected. Thus, FH correctly adapts to the wireless link's speed.

Another function of the BS is to inspect the regular ACKs returning from MH and to track the cumulative sequence number, which is inserted as the ACK number into the packets from BS to FH, hence the *unaltered end-to-end ACK information* from MH is conveyed to FH correctly. The CLACK agent does not harm TCP's end–to–end semantics and does not introduce any "hard state" into the system, because a TCP flow can recover from a CLACK agent failure without problems.

3) The Sender on FH: The sender is based on TCP Reno with the following extensions. *First*, the sender uses the CLACKs from BS as an indication that a packet has past BS, left the channel and new data may be sent. A missing CLACK indicates path congestion from FH to BS and leads to congestion recovery. *Second*, the sender uses the ACKs as an indication that a packet has been successfully read by MH.

The highest seen CLACK number which exceeds the highest seen ACK number (referred to as "CLACK" and "ACK" for convenience) indicates a corrupted packet to FH: the packet has past BS but is not read by MH and must be recovered. For this purpose, the sender is complemented with a "Corruption Recovery algorithm" in the same way the fast "fast recovery algorithm" is added in TCP Reno, refer to [13].

For several reasons, one may tolerate a certain difference between CLACK and ACK values. However this is left to future work, see the remark on flow regulation in Section VI. In the simulations used for this paper, we enter Corruption Recovery when the CLACK exceeds ACK.

B. Corruption Recovery

The Corruption Recovery algorithm is entered when the sender detects that $CLACK > ACK$. During Corruption Recovery, the FH is paced by the CLACKs sent by BS and missing data is retransmitted by FH until it is acknowledged by MH (which is seen in the ACK value received at BS). After

[1]The *average* Transport Block (TB) corruption rate may be an adaptation goal in some WWAN technologies. Although the *average* TB corruption rate may be known, the TB size is subject to change with the path conditions, hence it is unknown from how many TBs an IP packet is formed.

 ISBN:978-1-61208-113-7

the data is successfully delivered to MH, FH leaves Corruption Recovery and returns to normal TCP Reno operation. In other words, Corruption Recovery ends when the ACK and CLACK values are again identical.

In our implementation, only the first missing packet is retransmitted, causing the sender to enter Corruption Recovery several times when more than a single packet is missing. Although we look forward to finding better strategies, even this simple approach yields promising results, as shown in Section V.

C. Congestion Control during Corruption Recovery

The number of retransmissions necessary to recover from packet corruption is unknown in advance. In case of a low packet corruption ratio, a single retransmission may be sufficient, but in case of severe packet corruption, the number of required transmission attempts until eventual packet delivery may be arbitrarily high. Nevertheless, the following principles should be obeyed in packet retransmission.

1) The FH must not sent packets at higher speed than these can be conveyed to MH.
2) Packets sent by FH should not lead to path congestion.
3) Data retransmission must continue until the missing data is eventually delivered and acknowledged, unless otherwise agreed by the user or the application.

The fundamental idea of our approach is to employ the well-proven self clocking and congestion control strategy from the "congavoid paper" [4]. During Corruption Recovery, the retransmitted data and the corresponding CLACKs make up the "data in transit" in the sense of [4]. As long as the path from FH to BS remains uncongested, FH will be correctly paced by CLACKs received from BS and continue necessary retransmissions until FH receives the missing ACKs.

Unfortunately, we cannot rely upon the original TCP sliding window scheme because CLACK numbers sent by BS are not advanced by retransmissions. However, we can adapt the congestion control from [4] if we count *packets* instead of *bytes*. Without loss of generality, we can assume equally sized recovery packets.[2] Hence, the sending window's size corresponds to a maximum number of unacknowledged packets in the path. Let's denote the number of allowed packets in the path with as N.

In addition, we introduce a concept of recovery *rounds*. During a round, up to N packets can be injected into the path. When N packets are injected into the network, a new round is started. All packets belonging to the same round are marked by a common *timestamp* (refer to [19]), which is reflected by BS in the CLACKs. The timestamp can be the sending time of a round's first packet.

All packets sent in a round must be acknowledged within the RTO period. In other terms: For every round, the number of CLACKs with the according timestamp is counted. When the last packet was injected into the path, the last CLACK must reach FH within the RTO period. When the number of seen CLACKs is less than N, this is recognized as an indication of path congestion between FH and BS.

[2]In a very simple way, this can be achieved by retransmitting only the first missing packet as mentioned in Section III-B.

IV. Simulation Environment

For testing the performance of the CLACK agent and its impact on TCP, we have designed a Java-based discrete event simulator[3] which implements a simplified TCP Reno according to RFC 5681 [13]. Without loss of generality, we make the following assumptions and simplifications:

- We focus on simplex data flows with FH as a TCP sender and MH as the receiver.
- *Each* TCP packet is acknowledged by a "pure ACK", i.e., an empty ACK packet that contains no data.
- We focus on the "ESTABLISHED" state of a TCP flow, hence the startup and termination phases are omitted.
- A fixed TCP packet size is used in order to avoid a possible silly window syndrome and to facilitate a simple version of Nagle's algorithm.
- In the simulation setup for this paper, there is no congestion between FH and BS.

As shown in Figure 1, our simulation setup consists of an FH that is connected to a router by a full duplex 10 MBit/s link with a propagation delay of 3 ns. The router is connected to BS with a full duplex 10 MBit/s link and a propagation delay of *intentionally* 200 ms, because we are particularly interested in how the recovery algorithm works in presence of an unusual large propagation delay, which is rarely seen in terrestrial networks unless it is caused by WWANs with unreasonably high persistence in packet retransmission.

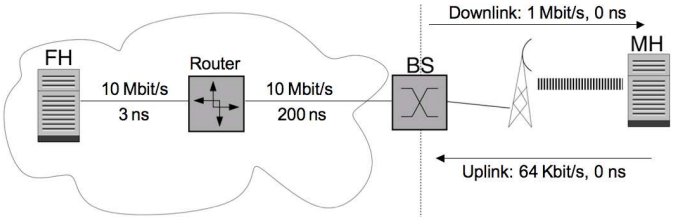


Fig. 4. Simulation setup.

The link between BS and MH is assumed to be a packet-switched WWAN, such as UMTS, GPRS or HSDPA. We assume a "Stop-and-Wait" protocol for the wireless link [10] and a reasonable limit for the number of transmission attempts per packet. As a "best current practice" recommendation, this limit is set to a maximum of *three* transmission attempts. The link between BS and MH roughly follows the HSDPA model: The *downlink* throughput is fixed at 1 MBit/s, the *uplink* throughput is 64000 Bit/s.[4] We performed simulations for several packet corruption rates in downlink direction. The

[3]Available at: http://detlef-bosau.de/index.php?select=tinysim

[4]We *do not* consider any propagation delay but used a fixed throughput, which leads to an appropriate serialization delay. Ideally, a WWAN interface should be modeled by its service time and SDU corruption rate. In our simulation, the packet sizes are fixed, so fixed service times and fixed throughput are equivalent in our case.

 ISBN:978-1-61208-113-7

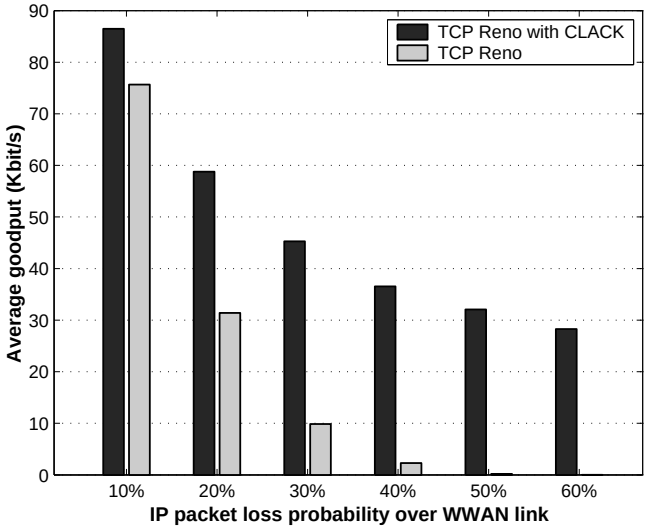


Fig. 5. Average goodput – standard TCP Reno vs. CLACK extensions.

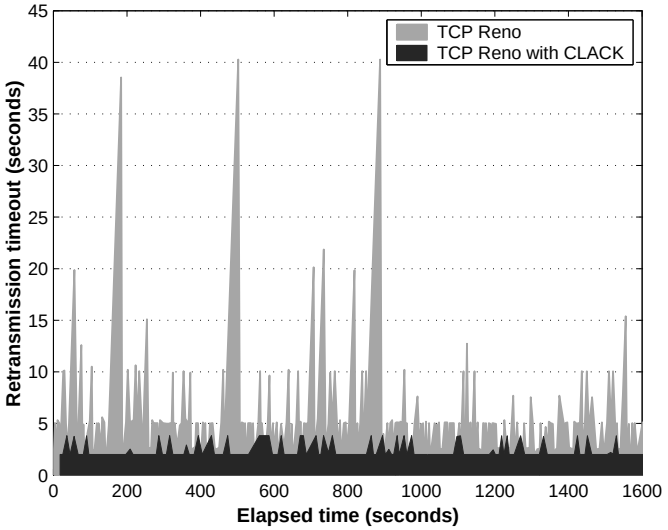


Fig. 6. Retransmission timeouts – standard TCP Reno vs. CLACK extensions (IP packet loss on WWAN link is set to 20%).

packet corruption rate in uplink direction is set to a fix value of 0.0. This simplification reflects the situation in HSDPA networks, where an extremely robust channel coding is used in uplink direction, at the expense of a quite low throughput.

WWANs typically employ some radio link protocol [20][10] which may convey a given packet in a certain *service time* and with a certain *corruption probability*. Actually, WWANs employ Stop-and-Wait protocols (refer to the discussion in [10]), because the wireless path's storage capacity is extremely small and a sliding window protocol would cause unreasonable overhead without providing significant benefit.

Protocols with "selective recovery" or sliding window mechanisms rarely make sense over WWAN links, because the link's storage capacity is typically extremely low and a minor throughput increase will hardly outweigh the overhead caused by a selective recovery protocol[10].

In the proposed protocol, the outgoing interface queue BS to FH is seen as a part of the "Internet part" of the path, hence the wireless network is free of congestion and any packet loss in this area is considered due to corruption. In addition, we assume *packet duplication* to be negligible.

V. Results

In this section, we discuss simulation results that show the performance of the CLACK approach. For the purpose of this paper it is not necessary to introduce a particular WWAN model. A WWAN link can be regarded as a link with constant bit rate and varying packet loss probability. For adaptive technologies, for instance HSDPA, the bit rate will change as well. However, this is beyond the scope of this paper and belongs to the discussion of RTT, refer to Section VI.

A. Average goodput

For the presence of various IP packet loss probabilities on the WWAN link, Figure 5 depicts the average goodput of a pure TCP Reno connection compared to a TCP Reno connection with a CLACK agent installed on the WWAN's base station. While TCP Reno becomes practically useless at error rates above 40%, the CLACK approach is able to maintain a low but steady data flow.

At a corruption error rate of 0% (not shown in Figure 5), both methods achieve an identical average data goodput of 954 Kbit/s, which is close to the 1 Mbit/s throughput of the error-free wireless link. Because our focus is on corruption recovery, we intentionally have no congestion in the wireline part of the path. Hence, there is no "congestion sawtooth" but only the TCP startup phase and the packet headers, which limits the goodput. In the absence of corruption errors, the Corruption Recovery mechanism introduced in Section III-B is never invoked. In other words, the CLACK agent has no negative impact on standard TCP Reno behavior.

B. Retransmission timeouts

The reason why the CLACK agent achieves an improved goodput at extreme corruption rates is because it avoids exponential RTO backoff to happen for packet loss that is caused by corruption on the wireless link. An exponential backoff should only be triggered due to congestion, not due to corruption. Without our proposed mechanism, TCP has to detect packet corruption by timeouts. In case of several sending attempts being necessary, these timeouts may grow significantly large due to the backoff algorithm, leading to phases without any data being delivered.

Figure 6 shows a timeline of two TCP Reno connections between FH and MH, one with a CLACK agent on the base station. In this experiment, the corruption probability was set to 20%, which may already cause timeouts of up to 40 seconds; leading to certainly noticeable pauses on the application layer. At the same time, the RTO values of the CLACK-enabled TCP Reno connection never exceeds 3.8 seconds.

For wireless corruption rates larger than 20%, the CLACK

 ISBN:978-1-61208-113-7

TABLE I
MAXIMUM RTO VALUES OBSERVED [SECONDS]

Corruption rate	10%	20%	30%	40%	50%	60%
TCP Reno	10	40	1310	10300	166000	$2 * 10^{12}$
CLACK	3.86	3.82	3.82	3.82	3.82	3.82

approach manages to maintain a stable, low RTO value, while a pure TCP Reno connection may experience RTOs in the order of hours if the loss rate exceeds 30%. Table I shows how immensely the RTO grows due to corruption under ordinary TCP Reno operation and how the CLACK approach achieves very stable values.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the CLACK approach to solve the problems of loss differentiation and corruption recovery for the widespread scenario of wireless networks being used as access networks to the Internet. Our scheme builds on the fact that standard TCP senders rely on positive acknowledgements from the receiver for proper pacing and injection of new data onto the end-to-end path. However, high packet loss that is not caused by congestion on the wireline path, but by corruption on the wireless access link, causes TCP to wrongly adapt to the available path capacity.

The central idea of our approach consists in pacing the TCP sender with "clock acknowledgements" (CLACKs) from the wireless base station, where the end-to-end path can be logically split into a wired part that can suffer from congestion (but not from corruption), and a wireless part that can suffer from corruption (but not from congestion). With the introduction of a CLACK agent on the base station, TCP is prevented from misinterpreting packet loss due to corruption as an indication of path congestion, which improves the utilization of the available capacity. In addition, our scheme allows a very fast recovery from packet corruption and therefore avoids extensive RTO growth and annoying pauses in TCP connections.

Although being a "middlebox approach", the CLACK agent *does not harm TCP's end-to-end semantics* but offers a supportive means to accommodate TCP to the challenges of mobile and pervasive networking. Additionally, our solution is incrementally deployable and compatible with standard TCP endpoints.

In our future work, we first will overcome the simplifications made in our system model and simulation setup in Section IV, including the following tasks:

- We will extend our work to bidirectional flows. So far, only a sender on FH can take advantage of the CLACK mechanism. However, the problem of loss differentiation and recovery also exists for a sender on MH. A particular task for bidirectional flows is to achieve proper pacing in *both* directions, allowing the bottleneck to reside in the wireline and the wireless parts of the path.
- A simplified recovery scheme is used in this paper. Particularly, the recovery from several lost packets could be done in a more sophisticated way.
- We consider a CLACK-based flow regulation mechanism which makes RTT appear quasi-stationary to the sender as a remedy for spurious timeouts.
- The current solution does not yet support delayed acknowledgments [21].

REFERENCES

[1] L. Zhang, "Why tcp timers don't work well," in *Proceedings of SIGCOMM*, 1986.

[2] R. Jain, "Divergenge of timeout algorithms for packet retransmissions," in *Proceedings of the Fifth Annual International Phoenix Conference on Computers and Communications, Scottsdale, AZ (USA)*, March 1986, pp. 174–179.

[3] S. W. Edge, "An adaptive timeout algorithm for retransmission across a packet switching network," in *Proceedings of the ACM SIGCOMM*, 1984, pp. 248–255.

[4] V. Jacobson and M. J. Karels, "Congestion Avoidance and Control," *ACM Computer Communication Review; Proceedings of the Sigcomm '88 Symposium in Stanford, CA, August, 1988*, vol. 18, 4, pp. 314–329, 1988.

[5] R. Ludwig, "Eliminating Inefficient Cross–Layer Interactions in Wireless Networking," Ph.D. dissertation, Aachen University of Technology, Aachen, Germany, April 2000.

[6] B. Ford and J. Iyengar, "Breaking Up the Transport Logjam," in *Proceedings of 7th Workshop on Hot Topics in Networks (HotNets-VII)*, October 2008.

[7] S. Fahmy, V. Prabhakar, S. R. Avasarala, and O. Younis, "TCP over wireless links: Mechanisms and implications," Purdue University, Tech. Rep. Technical Report CSD-TR-03-004, 2003.

[8] Y. Bai, A. T. Ogielski, and G. Wu, "Interactions of tcp and radio link arq protocol," in *In Proceedings of the IEEE VTC-Fall*, Amsterdam, The Netherlands, September 1999, pp. 1710–1714.

[9] H. Balakrishnan, "Challenges to reliable data transport over heterogeneous wireless networks," Ph.D. dissertation, University of California at Berkeley Department of Electrical Engineering and Computer Sciences, 1998.

[10] G. Fairhurst and L. Wood, "Advice to link designers on link Automatic Repeat reQuest (ARQ)," IETF RFC 3366, August 2002.

[11] J. Border, M. Kojo, J. Griner, G. Montenegro, and Z. Shelby, "Performance enhancing proxies intended to mitigate link-related degradations," IETF RFC 3135, June 2001.

[12] J. H. Saltzer, D. P. Reed, and D. D. Clark, "End–to–end arguments in system design," *ACM Transactions in Computer Systems*, vol. 2, no. 4, pp. 277–288, November 1984.

[13] M. Allman, V. Paxson, and E. Blanton, "TCP Congestion Control," IETF RFC 5681, September 2009.

[14] V. Paxson and M. Allman, "Computing TCP's Retransmission Timer," IETF RFC 2988, November 2000.

[15] J. Lacan and E. Lochin, "On-the-Fly Coding to Enable Full Reliability Without Retransmission," *Technical Report, ISAE, LAAS-CNRS, France*, 2008.

[16] J. Liu, I. Matta, and M. Crovella, "End–to–end inference of loss nature in a hybrid wired/wireless environment," in *Proceedings of WiOpt'03: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, 2003. [Online]. Available: citeseer.nj.nec.com/589180.html

[17] W. M. Eddy, S. Ostermann, and M. Allman, "New techniques for making transport protocols robust to corruption-based loss," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 5, pp. 75–88, 2004.

[18] C. Casetti, M. Gerla, S. Mascolo, M. Y. Sanadidi, and R. Wang, "Tcp westwood: Bandwidth estimation for enhanced transport over wireless links," in *Proceedings of ACM Mobicom 2001*, Rome, Italy, July 16–21 2001, pp. 287–297.

[19] V. Jacobson, R. Braden, and D. Bormann, "TCP Extensions for High Performance," IETF RFC 1323, May 1992.

[20] 3rd Generation Partnership Project 2 (3GPP2), "Data Service Options for Spread Spectrum Systems: Radio Link Protocol Type 3," 3GPP2 Document 3GPP2 C.S0017-010-A, June 2004.

[21] R. Braden, "Requirements for internet hosts – communication layers," IETF RFC 1122, October 1989.

 ISBN:978-1-61208-113-7

By Use of Frequency Diversity and High Priority in Wireless Packet Retransmissions

Xiaoyan Liu, and Huiling Zhu
School of Engineering and Digital Arts, University of Kent, Canterbury, Kent, CT2 7NT, United Kingdom
{xl36, h.zhu}@kent.ac.uk

Abstract – In orthogonal frequency division multiple access (OFDMA) systems, the delay of packets has a great impact on quality of service (QoS), especially for real-time transmission. The basic concept of resource allocation in an OFDMA system is to allocate a subcarrier to the user with the best channel condition in that subcarrier. However, when a retransmission technique is used, the packet delay of successfully transmitting failed packets may be quite large in conventional retransmission schemes. In this paper, the packet combining technique in the downlink OFDMA system is introduced in conjunction with frequency diversity to enhance the reliability of retransmissions and thus reduce the large time delay caused by retransmissions. The novel retransmission technique aims to reduce the maximum number of retransmissions for failed packets, while maximizing the throughput. Bit error rate (BER) and transmit power are constraints in the throughput maximization formulation. A suboptimal algorithm is proposed, in which the retransmitted packets have higher priorities to be given resources than new packets. At the receiver, failed and retransmitted packets are combined before detection by using maximal ratio combining (MRC). It is shown that the proposed retransmission scheme can reduce the maximum packet delay significantly, while maintaining the maximum throughput of the conventional retransmission schemes.

***Index Terms* – Orthogonal frequency division multiple access (OFDMA), real-time transmission, resource allocation, retransmission techniques, maximal ratio combining (MRC).**

I. INTRODUCTION

Future wireless communication services need to be provided with high data rate as well as guaranteed quality of service (QoS) such as delay and packet error rate (PER). Orthogonal frequency division multiple access (OFDMA) is a promising technique achieving high data rates for the next generation wireless systems. Adaptive resource allocation, which allocates resources such as modulation, power and number of subcarriers etc., adaptively to different users according to channel conditions, can utilize the radio resource efficiently due to the time-varying nature of the wireless channel. The nature of multiple subcarriers in OFDMA system leads to the possibility of adaptively choosing which subcarriers to be used among users, and at what rate and power to transmit on each subcarrier in each time slot.

Unlike packets in non-real-time applications (e.g. file transfer), which is considered delay-tolerant, packets in real-time applications are required to be constantly and correctly received by an end user. In real-time services, packets are often generated at regular intervals and required to be received with a delay constraint. Each packet has a delay deadline by which it must reach its destination; otherwise, it will be discarded. For example, in a full- motion video application [1], 30 frames are generated every second, and each frame must be delivered to the destination within a time delay to avoid any observable jitter. Therefore, transmission delay is an important problem in high data rate wireless communications especially for real-time traffic. When retransmission techniques are used, the maximum number of retransmissions for failed packets determines transmission delay. As known, automatic repeat request (ARQ) schemes are effective to recover non-real-time data corrupted by channel errors. However, with the high data rate in modern wireless networks, ARQ schemes are also favoured in real-time traffic [2]. Due to the QoS requirement of maximum packet delay caused by retransmissions of failed packets, in this paper a novel retransmission mechanism is proposed to reduce the delay introduced by retransmissions by allocating resources to retransmitted packets with higher priority than new packets.

In conventional resource allocation [3]−[5] in the OFDMA, retransmission techniques have not been considered for real-time traffic. Even if all retransmitted packets are guaranteed to be allocated subcarriers not being the best subcarriers, there is still a possibility that the retransmitted packets are received incorrectly. Thus, the maximum transmission delay in traditional resource allocation is long. In this paper, effective resource allocation is investigated when retransmission techniques are considered. The aim is to reduce the maximum number of retransmissions, so that the maximum transmission delay will be reduced. In other words, a new optimization problem is formulated to reduce the maximum packet delay by giving resources with a priority to retransmitted packets for transmission while maximizing the throughput in the downlink OFDMA system. In order to achieve high retransmission reliability, low density parity check (LDPC) [6]–[7], an advanced channel coding, is adopted since the constituent codes in the LDPC have parity check relationships and extremely good performance. Packet combining technique using maximal ratio combining (MRC) has also been adopted. In addition, the retransmitted packet will be transmitted through a subchannel which is different from the failed transmission to achieve frequency diversity.

The rest of the paper is organized as follows. The next section represents system model and is followed by Section III about the description of problem formulation. In Section IV, a suboptimal solution is proposed. Simulation results for different retransmission schemes and comparison among different allocation algorithms are given in Section V. Finally, this paper is concluded in Section VI.

 ISBN:978-1-61208-113-7

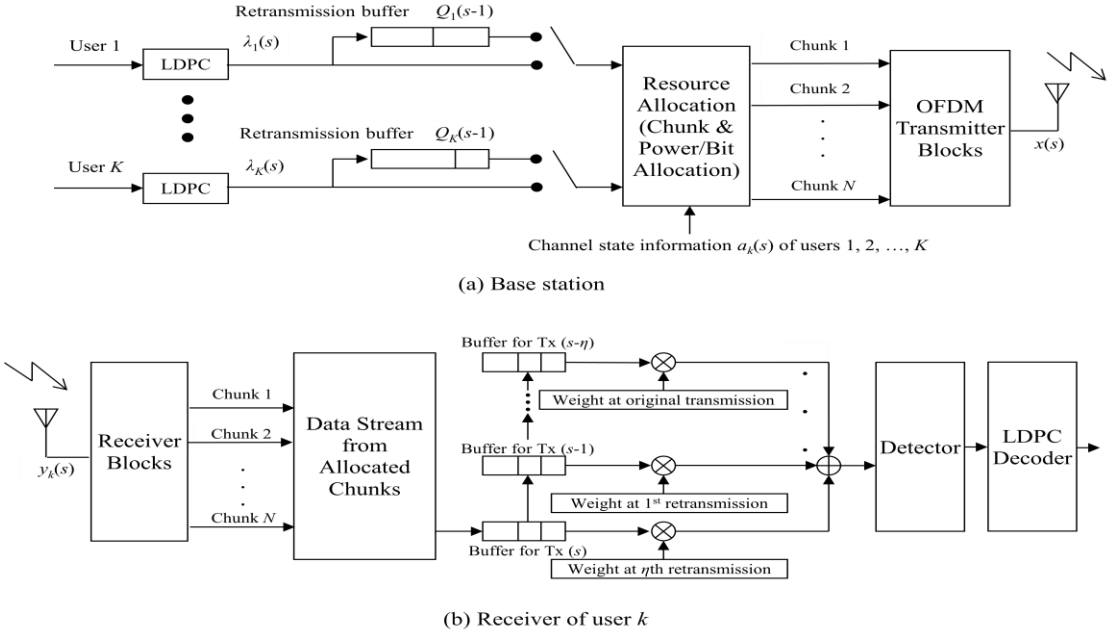


Fig. 1 Downlink multiuser OFDMA block diagram

II. SYSTEM MODEL

The downlink OFDMA system with N_{total} subcarriers and K users is shown in Fig. 1. In the base station (BS), resource is allocated periodically. An allocation period, denoted by τ, is equivalent to a downlink frame which includes several OFDM symbols. It is assumed that the channel condition is unchanged during one allocation period, i.e. one frame, so that the channel condition is the same for all the τ/T_s OFDM symbols in the frame, where T_s is the symbol duration and τ/T_s is assumed to be an integer. The time interval $(s-1, s]\tau$, where s is an integer, denotes the sth allocation period.

At the receiver of user k, the received signal in period s is expressed as

$$y_k(s) = \alpha_k(s)x_k(s) + v_k(s), \quad k = 1,2,\cdots,K \tag{1}$$

where $\alpha_k(s)$, called channel fading factor, is the magnitude of the frequency response of the channel fading of user k, which is independent from that of other users, $x_k(s)$ is the transmitted signal of user k, and $v_k(s)$ is a zero-mean additive white Gaussian noise (AWGN) of user k with bandwidth B and double-sided power spectral density $N_0/2$.

II A. Transmitter

The dynamic channel allocation is carried out on a packet-by-packet basis. As shown in Fig. 1, when a packet is ready for downlink transmission, it is numbered and stored in the input queue of the transmitter. After its transmission, the packet is saved in the retransmission buffer of the transmitter. When a positive acknowledgement (ACK) for a packet in the retransmission buffer is received, the packet is released from the retransmission buffer. If the receiver detects any error and feedbacks a negative acknowledgement (NAK), the transmitter will resend the failed packet. Any new packets from all users are encoded by the LDPC encoders, are allocated subcarriers and power, and transmitted together with packets to be retransmitted. The process is repeated until the packets are correctly received.

In the BS, the traffic source is considered greedy, which means that the system transmits as many packets as it can and there are always packets waiting to be transmitted. The BS allocates subcarriers and power to packets from all users according to their channel qualities, which may vary from frame to frame.

With the objective of minimizing the delay of retransmitted packets, all the retransmitted packets in queue, which are received incorrectly, will be guaranteed to be sent out during each allocation period (i.e. a frame). If a newly arrived packet is allocated with resources for its transmission, the packet will be encoded by the LDPC encoder before being sent to the transmitter block. By denoting the number of retransmitted packets failed in the $(s-1)$th period from user k as $Q_k(s-1)$, the number of newly arrived packets to be transmitted in the sth period from user k as $\lambda_k(s)$ and the number of acknowledged packets from user k in the sth period as $q_k(s)$, at the end of the sth period, the number of failed packets to be retransmitted for user k in the next transmission is given by

$$Q_k(s) = Q_k(s-1) + \lambda_k(s) - q_k(s), \quad k = 1,2,\cdots,K. \tag{2}$$

1) Subchannel Grouping

In this paper, the chunk allocation method [1] is used in allocating resources, where the number of bits and power are assumed to be the same for all the subcarriers within one chunk. In each allocation period, if a chunk is allocated to a

[1] In order to reduce the overhead and complexity of subcarrier allocation, the correlation between adjacent subchannels in the OFDMA system can be considered. Since the condition of a subchannel and its adjacent subchannels are quite similar, properly grouping a set of contiguous OFDM subchannels into one chunk and allocating the downlink resource chunk by chunk to users can make the subcarrier allocation simple while approaching the downlink throughput close to that of subcarrier-based allocation which allocates spectrum to users subcarrier by subcarrier [8].

user to transmit its newly arrived data, channel coding is exploited across the subcarriers in each chunk and the length of the new packet to be transmitted is adapted in order to be filled into one entire chunk. It is assumed that every m neighbouring subcarriers are grouped into one chunk. By grouping all subcarriers into N chunks, the computational complexity of resource allocation can be reduced approximately by $m = N_{total}/N$ times. To maintain transmission reliability while achieving high system throughput, the bit error rate (BER) constraint is given. Adaptive multi-level quadrature amplitude modulation (M−QAM) is adopted under this BER constraint. The adaptive modulation level M takes values from the set $\mathbf{M} = \{0, 4, 16, 64, 256\}$ and the corresponding data rate (bits per symbol) b takes values from the set $\mathbf{b} = \{0, 2, 4, 6, 8\}$. If the allocated bits of user k on each subcarrier over chunk n is $b_{k,n}$, the total bits for user k transmitted on this chunk is $mb_{k,n}$, and the bits of the new packet equals

$$L_{k,n} = mb_{k,n} \frac{\tau}{T_s}. \quad (3)$$

On the other hand, for a failed packet, the retransmitted signal will be combined at the receiver with that in the first transmission to achieve frequency diversity. In order to explore the frequency diversity, the retransmitted packet will be allocated by a chunk which is different from the previous one for transmission, and the bits of the retransmitted packet is the same as the original transmitted packet. Therefore, the bits allocated for retransmitted packets are set by the bits of the original transmission, and accordingly the power is allocated adaptively for the retransmission to guarantee the bits allocation.

2) Rate-Adaptive LDPC-Coded Modulation

In this system, similar to [9]−[12], the traffic source is encoded with $|\mathbf{M}|$ LDPC codes, where $|\mathbf{M}|$ denotes the size of $\mathbf{M}$, followed by Gray code mappings to $|\mathbf{M}|$ M−QAM constellations. When the M−QAM constellation is derived by the resource allocation in a frame, an LDPC code is constructed such that the bits of all coded packets can be fitted into τ/T_s OFDM symbols in the corresponding constellation. The length is calculated according to (3). The encoder and decoder hence have a given set of $|\mathbf{M}|$ LDPC generator matrices and corresponding parity check matrices, respectively. If chunk n is allocated to user k with code rate $R_{k,n}$, a source vector consisting of $R_{k,n}L_{k,n}$ information bits and $(1-R_{k,n})L_{k,n}$ parity check bits are encoded into the binary codeword of size $L_{k,n}$. The number of parity check bits to be transmitted in one of OFDM symbols in the frame is then equal to $(1-R_{k,n})mb_{k,n}$, whereas the number of information bits in each OFDM symbol is $R_{k,n}mb_{k,n}$. The number of information bits in all subchannels, which equals the system throughput (bits per OFDM symbol), is $R_{k,n}mb_{k,n}N$. This is carried out by letting the QAM modulator map to a transmitted signal vector of length τ/T_s, by taking τ/T_s distinct sets of $mb_{k,n}$ bits and mapping each such set into a complex OFDM symbol.

As greedy traffic is assumed, one packet occupies one chunk for its transmission and from (3), the packet length is determined by bits allocated to that chunk. In another word, the number of bits carried on the chunk is determined by the modulation level employed for that chunk. For simple implementation, chase combining technique [13] is used, in which exactly the same information bits and parity bits are contained in every retransmission. Because all transmissions are identical, chase combining can be seen as additional repetition coding. One retransmitted packet is only allowed to be transmitted over one chunk rather than over many. Thus, in order to fill a retransmitted packet within a chunk, the number of bits on the chunk is predetermined by the failed packet's length or its previous modulation level, only if it is known that this chunk is assigned to retransmitted packets. Except for chunks allocated to retransmitted packets, the sizes of the other chunks are still determined by their bits allocation as in (3). In another word, same modulation level is employed for each retransmitted packet as in its previous transmission since they have the same packet length.

II B.Receiver

The modulated M−QAM symbol vector is then transmitted through a slowly varying flat fading channel. Such a vector will then be subjected to (almost) constant fading and the added channel noise v. As shown in Fig. 1, the data stream from the allocated chunks of user k is split through a chunk selection block. The demodulator uses maximum likelihood soft decision detection on the received data stream, while the subsequent decoder uses an iterative belief propagation decoding algorithm [14].

In the recovered data stream, if a packet is received incorrectly, it will be stored in the receiver buffer, and this packet must be retransmitted immediately in the next frame to guarantee the delay constraint of real-time services. Along with the previous received incorrect packet stored in the receiver buffer, the retransmitted signals could be combined by implementing MRC. One major issue is that sufficient chunks are assumed to guarantee the transmission of failed packets, and retransmitted packets may be transmitted in different chunks from previous transmission, which depends on the chunk condition. In our proposed allocation scheme, failed packets have higher priority to be allocated the best chunks of their users. By implementing this sort of frequency diversity technique, high transmission success could be achieved.

The buffers in the receiver are used to save failed packets together with the channel state information from the user. The received signals from different transmission slots of the same packet are saved separately in buffers for the corresponding transmission slot, as in Fig. 1. Suppose a packet is received correctly after η retransmissions, then at the receiver side η buffers are used to save this packet from allocated chunks of user k in transmission slot $s-\eta$, $s-1$, …, and s. Along with the weights, proportional to channel state information, of these transmissions, the receiver uses the MRC to combine the

received signal with the signals received from previous transmissions. The detector is the demodulator with log-likelihood ratio (LLR) output and is followed by a LDPC decoder with hard decision, where the number of iterations to be performed for decoding one codeword is given.

III. PROBLEM FORMULATION FOR RESOURCE ALLOCATION

Here, optimizing the chunk and power allocation is considered under the power constraint with rate discretization, so as to reduce the maximum packet delay caused by retransmission while maximizing the throughput of new packets of all users. The retransmitted packets and new packets are treated differently in resource allocation by taking into account of the trade-off between maximizing system throughput and reducing packet delay. The optimal chunk assignment is considered for a given set of total available chunks defined as $\mathbf{I} = \{1,2,\cdots,N\}$. The current allocation period index s is omitted for some variables in the following for simple notation. The optimization problem can be mathematically formulated as

$$\max_{\{\Omega_k, p_{k,n}\}} \sum_{k=1}^{K} \sum_{n\in\Omega_k} b_{k,n} \tag{4}$$

subject to

$$\sum_{k=1}^{K} \sum_{n\in\Omega_k} m p_{k,n} \le P_T - \sum_{k=1}^{K} \sum_{n'\in\Omega'_k} m p_{k,n'} \tag{4.1}$$

$$p_{k,n} \ge 0 \text{ for all } k, n \tag{4.2}$$

$$b_{k,n'} = b_{k,n}(s-1), n' \in \mathbf{\Omega}'_k, n \in \mathbf{A}(s-1) \tag{4.3}$$

for all k

$$\left(\bigcup_{k=1}^{K} \mathbf{\Omega}_k\right) \cup \left(\bigcup_{k=1}^{K} \mathbf{\Omega}'_k\right) \subseteq \mathbf{I} \tag{4.4}$$

$$|\mathbf{\Omega}'_k| = Q_k(s-1) \text{ for all } k \tag{4.5}$$

where the prime, (’), denotes notation on retransmissions. $p_{k,n}$ is the power allocated to user k on each subcarrier in chunk n, and it is constant across the subcarriers in each chunk. Constraints (4.1) and (4.2) guarantee that allocated power on all chunks is non-negative and the total power does not exceed the total power constraint P_T. $\mathbf{\Omega}_k$ and $\mathbf{\Omega}'_k$ are the sets of chunks assigned to user k for its packets on first-time transmissions and retransmissions, respectively. In constraint (4.3), the assigned transmit bits per symbol of allocating one retransmitted packet of user k, equals its previous allocated bits $b_{k,n}(s-1)$, which is an integer, if the allocated chunk in previous allocation period is $n \in \mathbf{A}(s-1)$, where $\mathbf{A}(s-1)$ is the set of previous chunk allocation.[2] Its assigned transmit bits per symbol in this allocation period is $b_{k,n'\in\mathbf{\Omega}'_k}$. (4.3) also shows that the modulation level of the retransmitted packet must be the same as that in the previous transmission in order to combine retransmitted packet with previous failed transmitted packet(s) to achieve frequency diversity. Both new and retransmitted packets for all users are allocated by chunks from the set of total available chunks, **I**, as in (4.4). (4.5) assures that all retransmitted packets $Q_k(s-1)$ can be sent out by assigning sufficient number of chunks, $|\mathbf{\Omega}'_k|$ is the number of chunks for the retransmitted packets of user k.

Assuming that the retransmitted packets from every user have high priority to be transmitted, they are allocated sufficient chunks within the allocation period in order to guarantee the transmission of all the retransmitted packets. From (4.1), the remaining power from the consumption by retransmitted packets is allocated to new packets with the objective of maximizing the throughput. Thus, the optimization problem in (4) can be split into 2 steps. For the retransmitted packets of all users, since the bits to be transmitted are given, the problem turns to minimize the overall power consumption on retransmission.

After solving the above problem with all these constraints, a set of chunks occupied by retransmitted packets $\mathbf{\Omega}'_k$ for all k, $\mathbf{\Omega}'_k = \{n' | \rho'_{k,n'} = 1\}$ can be obtained. The set of chunks available for new packets $\mathbf{\Omega}_k$ is thus obtained as the rest of **I**. By assuming all new packets have the same BER requirement, $b_{k,n\in\mathbf{\Omega}_k}$ could be chosen for any user k according to the required received signal to noise ratio (SNR) and the SNR gap Γ [9], which is a constant related to this given BER requirement if the same modulation scheme is used.[3] The value of Γ depends on the capacity difference between practical system using adaptive modulation and Shannon capacity (Γ=1 (0dB)). $b_{k,n}$ can be expressed approximately as

$$b_{k,n} = \log_2\left(1 + \frac{p_{k,n}|\alpha_{k,n}|^2}{\Gamma N_0 B/N}\right), \text{ integer for all } k, n \tag{5}$$

where $\alpha_{k,n}$ is the channel fading factor for user k in chunk n. $b_{k,n}$ can be determined through the received SNR region given in Table I, conditioned on different values of BER constraint.

IV. PROPOSED ALLOCATION ALGORITHM

Firstly, let A_n, b_n and p_n denote the user index, number of bits and power allocated on chunk n, respectively. $H_{k,n} = |\alpha_{k,n}|^2/N_0(B/N)$ is defined as the channel to noise ratio for user k in chunk n. With the bits allocation $\{b_n(s-1)\}$ and chunk allocation $\{A_n(s-1)\}$ of retransmitted packets in previous resource allocation $n \in \mathbf{A}(s-1)$, which are registered in two separate registers embedded in the BS’s resource allocation block. In order to guarantee the success of retransmitted packets, each packet on retransmission has higher priority to choose its best chunk. The proposed allocation determines the number of bits and chooses ‘better’ chunks, which has higher channel to noise ratio, for retransmitted packets according to the algorithm shown below. Given the BER constraint, the number of bits $\{b_n\}$ and chunk conditions $\{\alpha_n\}$on allocated

[2] The value of each element A_n in **A** is the user index on subcarrier n.

[3] If no coding is considered, for M−QAM $\Gamma = -\ln(5BER)/1.6$, where BER is the BER requirement [11].

chunks for retransmitted packets, the power distribution $\{p_{n'}\}$ thus can be obtained from the transformation of (5). The rest of available chunks will be allocated to new packets according to the *Maximum-Capacity* allocation algorithm which is explained in the following paragraph.

In order to maximize system throughput for new packets, multiuser diversity is explored by allocating chunks to the user with the best channel condition on that chunk. In this algorithm, the transmit power is firstly assumed to be the same for subcarriers in one chunk and then an optimal power allocation algorithm is carried out among chunks to maximize throughput. The optimal transmit power adaptation is the well-known water-filling procedure [15]. In water-filling, more power is allocated to "better" chunks with higher SNR, so as to maximize the sum of data rates in all chunks.

The new packets are firstly allocated with chunks within the rest of available chunks according to their channel to noise ratio. The initial power distribution is then obtained by water-filling algorithm, so that corresponding initial bit rate can be calculated according to (5), which can be a continuous value. Because of the integer nature of bits per subcarrier per symbol, data rate calibration (as described in step 3-c below) has to be carried out to fit the existing modulation scheme. Finally, the transmit power allocated over its assigned chunk is recalculated accordingly. The algorithm can be described as:

1) Initialization

a) Set $b_n = 0$, $A_n = 0$, $p_n = 0$ for $n = 1, 2, \cdots, N$.

2) For Retransmitted Packets:

for i = 1 to number of retransmitted packets

a) find n' satisfying $|H_{k,n'}| \geq |H_{k,n}|$ for all n;

b) let $b_{n'} = b_n(s-1)$, $A_{n'} = A_n(s-1)$;

c) calculate $p_{n'}$ according to (5).

end

3) For New Packets on $n \in \{A_n = 0\}$:

for i = 1 to N – number of retransmitted packets

a) find k' satisfying $|H_{k',n}| \geq |H_{k,n}|$ for all k;

b) let $A_n = k'$, use water-filling algorithm to obtain $p_{k',n}$ among the rest power $P_T - \Sigma\, p_n$;

c) calculate $b_{k',n}$ according to (5), let $b_n = \lfloor b_{k',n} \rfloor_{\mathbf{M}}$, where $\lfloor x \rfloor_{\mathbf{M}}$ denotes the largest number in the set of **M** less than or equal to x.

d) calculate p_n according to (5).

end

Then, $\mathbf{A} = \{A_{n'}\} \cup \{A_n\}$ is the chunk allocation, $\mathbf{B} = \{b_{n'}\} \cup \{b_n\}$ is the bit allocation, $\mathbf{P} = \{p_{n'}\} \cup \{p_n\}$ is the power allocation.

V. SIMULATION RESULTS

In this section, the performance of the OFDMA system with the proposed resource allocation algorithm is investigated. It is assumed that an OFDMA system has 512 subcarriers (N_{total} = 512), the number of subcarriers per chunk is 16 (m = 16), the number of users is eight (K = 8) and symbol duration T_s = 20μs. A frequency-selective Rayleigh fading channel is assumed, which is a 17-path channel with exponential power delay profile and root mean square (RMS) delay spread of 0.5μs for each path. Doppler frequency is 1Hz. The channel fading factor $\{\alpha_{k,n}^2\}_{n = 1, 2, \ldots, N}$ of each chunk is Rayleigh distributed with $E\{|\alpha_{k,n}|^2\} = 1$. The BER constraint is 10^{-3} with 1/2 rate LDPC coding, where the parity check matrix is even in both rows and columns, length-four cycle is eliminated, and number of ones per column is 3. Greedy traffic source is assumed. That is, the traffic source of user k generates packets at the rate depending on allocated bits per symbol $\{b_{k,n}\}$; and each packet contains flexible length of information bits accordingly. Each frame consists of 6 OFDM symbols. The number of packets the system could allocate within one frame is N = 32, since each chunk carries only one packet. 100 frames are chosen for each simulation run.

Given the target BER, two allocation schemes are investigated by taking packet retransmissions into account. One is the conventional *Maximum-Capacity* scheme, by allocating each chunk to the user with the best chunk condition for that chunk. The other one is our proposed allocation scheme with packet combining, where all the packets on retransmissions are allocated with their best chunks, and assigned same modulation as in original transmission; new packets are allocated among the remaining chunks based on the conventional *Maximum-Capacity* allocation scheme. Packet combining is also adopted at the receiver to improve the performance. Each retransmitted packet is sent out immediately for both schemes as long as its user is allocated sufficient chunks, otherwise it will wait in the retransmission buffer until the next frame. The following metrics are used to evaluate the performance through the simulation:

- *Packet Error Rate*, defined as the total number of error packets divided by the total number of packets sent.
- *Average Packet Delay*, defined as the average number of retransmissions of each successful packet.
- *Cumulative Distribution Function of Number of Retransmissions*, where the probability distribution of each number of retransmissions is defined as the ratio of amount of each number of retransmissions to total number of successful packets within the simulation run.

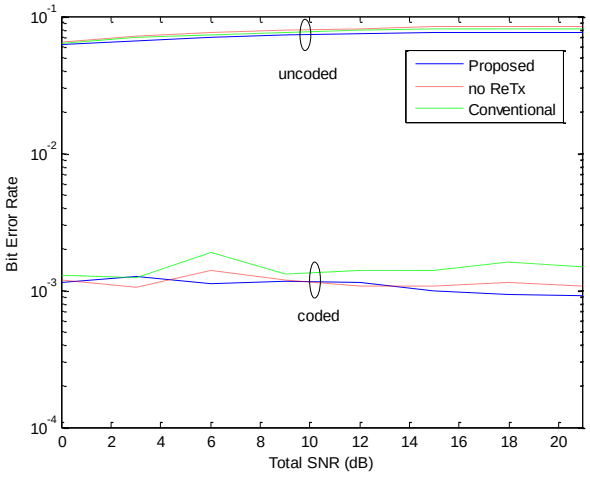


Fig. 2 BER versus total transmit SNR

Fig. 2 shows the BER performance for both the uncoded and coded systems when SNR takes values from 0dB to 25dB. In this figure, *Proposed* denotes the proposed allocation scheme, *Conventional* denotes the conventional *Maximum-Capacity* allocation scheme, and *noReTx* denotes the conventional allocation scheme without retransmission. The notations are also used in the following simulation results. Given the target BER of the LDPC coded to be 10^{-3}, due to the discrete rate adaptation and constant power restriction, the instantaneous BER fluctuates very slightly around the target BER, i.e. 10^{-3}, as SNR varies. The average BER ranges from $0.9594*10^{-3}$ to $1.913*10^{-3}$ at the target BER 10^{-3}. The BER performance is almost the same for the proposed scheme and conventional one when the same BER constraint is given. The corresponding uncoded average BER increases monotonically as the SNR increases and is within an order of magnitude for all the SNR values. The uncoded BER results of the proposed and the conventional schemes are quite close, while both go up slightly as SNR increases due to the impact of the PER. It shows BER constraint is guaranteed under both schemes.

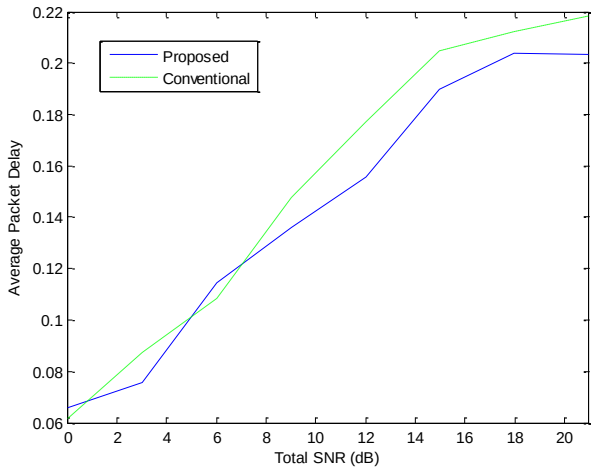


Fig. 3 Average packet delay versus total transmit SNR

The average packet delay is shown for the two allocation schemes in Fig. 3. It is assumed that the transmission of a packet will not be completed until the end of a frame, i.e. each transmission period is a frame. Therefore, the time delay equals the value of delay multiplied by the frame period τ. The number of retransmissions equals packet errors because once a packet is received in error, a retransmission will be initiated; and the total number of packets sent approximates to that of the successful packets when PER is low. Thus, average packet delay has the similar trend as PER according to their definitions.

VI. CONCLUSION

In this paper, a novel adaptive chunk and power allocation scheme with packet retransmissions in the downlink OFDMA system is proposed, which supports ARQ. In order to reduce the maximum number of retransmissions for failed packets while maximizing the throughput of new packets in the system, a new optimization is formulated under a target BER and a total transmission power constraint. Based on that the retransmitted packets have higher priorities to be allocated with resources, the proposed allocation scheme with the use of packet combining technique has better performances than the conventional scheme in terms of the maximum packet delay.

REFERENCES

[1] B. Melamed, "Modeling compressed full-motion video", *Proceedings of the 29th Conference on Winter Simulation*, pp. 1368−1374, 1997.

[2] J. Neckebroek, H. Bruneel, and M. Moeneclaey, "Application layer ARQ for protecting video packets over an indoor MIMO-OFDM link with correlated block fading," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 467–475, Apr. 2010.

[3] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[4] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[5] Y. J. Zhang, and K. B. Letaief, "Cross-layer adaptive resource management for wirless packet networks with OFDM signalling," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3244–3254, Nov. 2006.

[6] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, 2nd ed., Prentice Hall: Englewood Cliffs, NJ, 2004, Ch. 17, Ch. 22.

[7] M. Yang, W. E. Ryan, and Y. Li, "Design of efficiently encodable moderate-length high-rate irregular LDPC codes," *IEEE Trans. Commun.*, vol. 52, no. 4, pp. 564–571, Apr. 2004.

[8] H. Zhu, and J. Wang, "Chunk-based subcarrier allocation in OFDMA systems: part I: chunk allocation," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2734–2744, Sept. 2009.

[9] J. G. Proakis, *Digital Communications*, 4th ed., New York: McGraw-Hill, 2001.

[10] M.-S. Alouini, and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Kluwer J. Wireless Commun.* vol. 13, pp. 119−143, May 2000.

[11] S. T. Chung, and A. J. Goldsmith, "Degrees of Freedom in Adaptive Modulation: A Unified View," *IEEE Trans. Wireless Commun.*, vol. 49, no. 9, pp. 1561–1571, Sept. 2001.

[12] B. Xia and J. Wang, "Effect of channel estimation error on QAM systems with antenna diversity," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 481–488, Mar. 2005.

[13] N. Miki, H. Atarashi, K. Hiquchi, S. Abeta, and M. Sawahashi, "Experimental evaluation on effect of hybrid ARQ with packet combining in forward link for VSF-OFCDM broadband wireless access," *IEEE Proceedings on Personal, Indoor and Mobile Radio Commun.*, PIMRC 2003, vol. 1, pp. 360–365, Sept. 2003.

[14] R. Lucas, M. Fossorier, Y. Kou, and S. Lin, "Iterative Decoding of One-Step Majority Logic Decodable Codes Based on Belief Propagation," *IEEE Trans. Commun.*, 48 (6): 931–37, Jun. 2000.

[15] W. Yu, and J. M. Cioffi, "On constant power water-filling," *IEEE International Conference on Commun.*, ICC 2001, vol. 6, pp. 1665–1669, Jun. 2001.

Kernel Monitor of Transport Layer Developed for Android Working on Mobile Phone Terminals

Kaori Miki
Masato Oguchi
Department of Information Science
Ochanomizu University
2-1-1, Otsuka, Bunkyo-ku, Tokyo, Japan
kaori@ogl.is.ocha.ac.jp
oguchi@computer.org

Saneyasu Yamaguchi
Kogakuin University
1-24-2 Nishi-shinjuku, Shinjuku-ku, Tokyo, Japan
sane@cc.kogakuin.ac.jp

***Abstract*—In recent years, with the rapid growth of smart phone market, Android is drawing an attention as software platform of embedded system, used as a personal digital assistance developed by Google. While Android is taken notice for its flexible development of application software and expansion of the system, we are interested in optimization and performance evaluation of network computing ability of Android. Because an embedded system like Android has architecture different from that of general-purpose PC, and due to the poor function of I/O interface, it is difficult to grasp what happens inside the embedded system precisely. Therefore, it is interesting to analyze the communication behavior of Android. In this paper, we have developed a Kernel Monitor tool suitable for an embedded system that is able to observe the behavior of kernel. We have applied this tool for the Transport layer of Android. We have shown that internal operation when an embedded system is communicating can be analyzed with our approach.**

***Keywords*-Android, Mobile Phone, Embedded system, Linux Kerne,**

I. INTRODUCTION

Recently, almost every user has own cellular phone. Moreover, it is not rare to have two or more cellular phones by a single user, and those phones are used in different ways depending on service and the usage. Previously, a cellular phone was used only for the voice call and text mail. However, since the transmission rate improves recently, it is applied to many functions including Internet access, distribution of music and animation, radio, television, IC card, and so on. Therefore, it is difficult to develop applications for a unique OS of each career due to its huge cost. Thus, commoditized OS for cellular phones has been desired. In this case, the basic part is shared as a platform, and original functions and services are developed individually. As a result, the efficient improvement of application development can be achieved. Moreover, since there is an advantage that it becomes easy for programmers to develop applications on it, the number of open software should be increased. Android [1] has been developed by Google for this purpose. Android is a software platform of an embedded system that works with portable devices.

Android is different from software used in current mobile phone terminals. There is no restriction for the application development because this is open source software. Applications can be executed on mobile phone equipped with Android regardless of the career or the device, and they can be highly customized. From these factors, the share of Android is increasing. While Android is drawing attention for flexible development of application software and expansion of the system, we are interested in Android as a system platform. In particular, since mobile phones, such as smart phones, become a leadingwe aim to optimize and evaluate performance of network computing ability of Android. Most of recent research works on Android concentrate only on applications, except [2] in which CPU load of Dalvik bytecode is investigated. Our research works are focused on the communication ability of Android platforms.

Because architecture of an embedded system is different from that of general-purpose PC, it is interesting to analyze its communication behavior. In particular, since the mobile phone such as smart phones becomes a leading part as a client terminal, to analyze their behavior is drawing attention in a variety of communication scenes. However, as for the behavior of the embedded system, the observation method is extremely limited due to the poor function of I/O interface. Because the resources of an embedded system is much less than that of general-purpose PC, the resource that can be spared for the monitoring and analysis is insufficient. Therefore, it is possibly considered having a substantial influence on the behavior of the system by its monitoring. From such a reason, it has been difficult to grasp exactly what happens in embedded systems during communications.

In this paper, since we have overcome the difficulty of an embedded system such as different interfaces and so on, we have developed a Kernel Monitor tool in Android, which has basically the same function of such a tool developed for general-purpose PC [3]. We have applied the tool in

 ISBN:978-1-61208-113-7

the Transport layer of Android. As a result, we show that the internal operation of an embedded system when it is communicating can be analyzed with this approach.

II. Development of Applications on Android

In this section, we explain about Architecture of Android and how to develop applications on it.

A. An Overview of Android

Table I
Architecture of Android

Application(Home,Telephone,Web)
Application Framework
Android Runtime Core Libraries, Dalvik VM Ware
Library
Linux Kernel 2.6

The architecture of Android is shown in Table 1. Android is constructed with Linux Kernel 2.6, and various components are added to its OS so as the platform to be composed. Because only the kernel part is adopted from Linux, it is possible to compose Android from various Linux packages.

Android Runtime, that is the application execution environment of original Android, is mounted on Linux kernel. The original virtual machine called Dalvik is installed in Android. This corresponds to Java Virtual Machine (JVM). Applications can be developed just suitable for Dalvik, because the application frameworks are provided on top of Dalvik for the execution of applications. Therefore the portability of Android is very good.

Android is different from other software implemented in current portable devices, and there is no restriction in the application development because this is open source software. Applications can be executed on mobile phone equipped with Android regardless of the career or the device, and they can be highly customized. Thus the load of the application development is considered to be reduced, and there is a flexible extendibility to another career and another model.

The communication is performed by using the protocol stack in Linux kernel. Thus, it is thought that the communication performance of Android is decided in this TCP implementation. Therefore, Transport layer in the kernel is highlighted and evaluated in this research work.

B. Cross Development

The cross development that uses a different computing environment is employed when applications of the embedded system is developed and executed. This is because mobile terminal's display is too small, and neither CPU performance nor memory capacity is enough. In Android, cross development is generally employed. The computer that develops chiefly is called a host environment, and Android terminal is called a target environment. The camera equipment and so on that the host environment does not usually include can be executed by the emulator inside hosts. Android is not suitable for development environment, because Android is an embedded system that has only limited commands compared with the case of general-purpose PC. Cross development can raise the efficiency of development.

Not only application development but also build of Android itself are executed on the form of the cross development. The Kernel Monitor introduced in this paper is also formed on the cross development, then implemented in the Android terminal.

III. Kernel Monitor

In this section, we explain Kernel Monitor which is our original system tool, and how to develop Kernel Monitor that works on Android terminal.

A. An Overview of Kernel Monitor

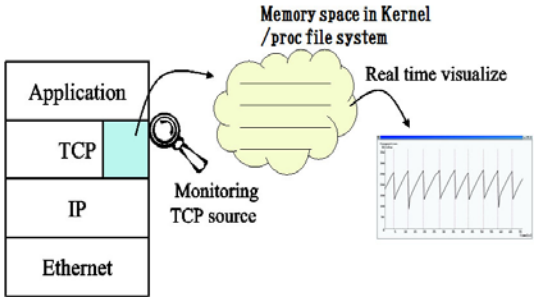


Figure 1. An Overview of Kernel Monitor

Kernel Monitor is a tool that can recode how the value of parameters in the kernel changed as a result by which part of the kernel code being executed at which time when communicating. An overview of Kernel Monitor is shown in Figure 1. We have inserted the monitor function in the Transmission Control Protocol(TCP) source code of the kernel, then recompiled the kernel, TCP parameters can be monitored as a result. Examples of what we can monitor with this tool are the value of Congestion Window (CWND) and various error events of communications (Local device congestion, duplicate Acknowledgment (ACK)/ Selective Acknowledgment Options (SACK), and Timeout). The behavior of kernel can be shown when it normally operates with the Kernel Monitor. In addition, when something wrong happens, it is possible to detect a specific problem and investigate what happens inside the kernel.

Kernel is special software different from other applications. It cannot accept normal debug methods, and therefore, it is difficult to observe the behavior of kernel during communications even in the case of general-purpose PC. However, in general-purpose PC, this problem has been solved by using Kernel Monitor [3]. In this paper, we

 ISBN:978-1-61208-113-7

have applied the Kernel Monitor to Android, an embedded system.

B. Development of Kernel Monitor for Android Terminals

Since the base of Android is Linux kernel, Android has a possibility to accept similar approach of general-purpose PC. However, as Android is an embedded system, it has a lot of different points from the case of general-purpose PC. For example, the amount of resources of Android terminals, such as storage and memory, are limited. Thus, the same approach as general-purpose PC may be impossible due to resource shortage. In addition, since the resource that can be spared for the operation analysis is insufficient, there is a possibility of having a substantial influence on the behavior of the system by the operation analysis itself. Android also need the cross compile for system and applications. OS is built in a special way. Moreover, it is required to use a special way to boot compiled OS on an Android mobile phone.

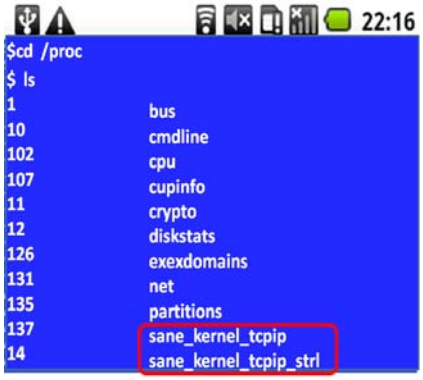


Figure 2. /proc file system

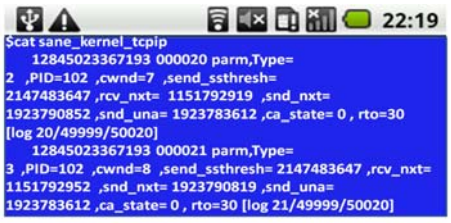


Figure 3. Log of Kernel Monitor

In this paper, we have overcome difficult problems peculiarly an embedded system's own, and developed Kernel Monitor that is the same as general-purpose PC basically. This Kernel Monitor has been implemented in Android OS codes, inserted into an Android mobile phone terminal, and we have confirmed that the Kernel Monitor runs on it. We show that it is possible to analyze the behavior of Android when it is communicating. Figure 2 and 3 are examples of captured display that Kernel Monitor is running on Android mobile phone. As a result, it becomes possible to show that Android's kernel behavior can be analyzed with almost the same way as the Kernel Monitor of general-purpose PC. Thus, we have realized the monitor tool in this environment.

As an example of monitoring, we have analyzed the relation between Congestion Window and throughput during the communication on Android. It is introduced in the following sections.

IV. Experimental System and Measuring Basic Performance of Android

In this section, measurement tool and experimental environment are shown, and the experimental way is introduced.

Table II
Experimental Environment

Android	Model number	AOSP on Sapphire(US)
	Firmware version	2.1-update1
	Baseband version	62.50S.20.17H_2.22.19.26I
	Kernel version	2.6.29-00481-ga8089eb-dirty
	Build number	aosp_sapphire_us-eng 2.1-update1 ERE27
server	OS	Fedora release 10 (Cambridge)
	CPU	CPU : Intel(R) Pentium(R) 4 CPU 3.00GHz
	Main Memory	1GB

Table II, Figure 4 and Figure 6 show the experimental environment. In our study, we have cross-compiled iperf-2.0.4 [4], and inserted it Android mobile phone. With this tool, we have evaluated the socket access case as a basic performance. Arm-2008q3 [5] is used as a cross compiler.

A. Android to Android Communication Throughput

Figure 4. Android to Android Communication

First, we have evaluated Android mobile phone's throughput with IEEE 802.11g Wireless Local Area Network (LAN) through Access Point (AP) to another Android mobile phone, as shown in Figure 4.

Performance of socket access using TCP and User Datagram Protocol (UDP) is shown in Figure 5.

TCP communication average throughput of Android to Android is 8.2 (Mbps), while that of UDP communication is 6.7 (Mbps), as shown in the Figure 5. According to this graph, performance of UDP access is lower than that of TCP in the case of Android mobile phone terminals. There is almost no packet loss in both cases. The causes of throughput degradation seems to exist at the sender-side. The packet can be sent out only on a constant rate even if the

 ISBN:978-1-61208-113-7

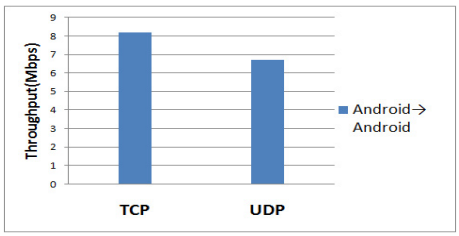


Figure 5. TCP thoughput of Android to Android Communication

bandwidth at the transmitting end is enlarged. Moreover, in UDP access, although we have not confirmed, performance of UDP access seems to be limited in the case of off-the-shelf devices. By way of comparison, we have evaluated Android-x86 that runs on x86 PC platform [6]. In this case, we have confirmed that throughput of UDP access is higher than that of TCP access.

B. Android Communication Performance of Remote Server Access

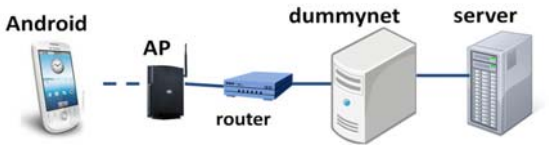


Figure 6. Android Communication Performance of Remote Server Access

We have evaluated Android communication performance of remote server access with dummynet (see Figure 6), which artificially generates delay. This supposes to access to a server on a remote place, which offers mobile cloud service.

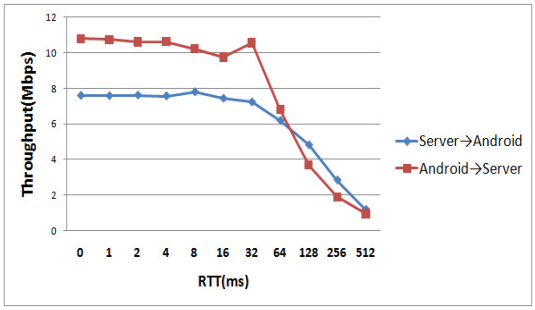


Figure 7. TCP Throughput between Server and Android Terminal in a Higher-Latency Environment

Figure 7 is a graph of throughput in which the horizontal (x-)axis is Round Trip Time (RTT) by dummynet. Better throughput is observed in the case that server is used as a receiver-side, because server's receive buffer is larger than that of Android terminal. However, in a higher latency environment, the performance in which server is receiver is declined. This is because Android cannot increase CWND enough in a higher-latency environment, and therefore CWND is running out. We have applied Kernel Monitor developed for Android, and analyzed CWND. It is saturated at 66: with this value CWND is not enough. Further details about Kernel Monitor is explained in the next section.

V. TCP TUNING AND APPLY KERNEL MONITOR TO ANDROID

We have tried to evaluate the effectiveness of changing Congestion Window control algorithm in TCP/ Internet Protocol (IP). We call it TCP Tuning in the rest of this paper. Two Congestion Window control algorithms are available for Android — Reno and Cubic (default). One of the most suitable Congestion Window control algorithm for mobile terminal is Westwood that tolerates packet loss, which cannot be applied to Android in this case. This depends on downloaded source code of Android when it is built.

A. Congestion Control Algorithms

A lot of congestion control algorithms are discussed in the literatures [7][8]. Reno is the basic algorithm. A wide variety of algorithm has been developed based on Reno. Reno detects congestion by packet loss and it regards transfer rate at the time as available bandwidth. For example, if three consecutive Duplicate ACKs are received, Reno regards it as occurring of packet loss and reduces CWND by half. Moreover, it increases CWND on receiving every ACK. Thus, Reno is an algorithm that increases the size of CWND gradually and drops it by detecting congestion. CUBIC [9][10] is an improved algorithms of Binary Increase Congestion Control (BIC). BIC is an algorithm that in normal TCP congestion control, linear search is executed for available bandwidth, on the other hand, binary search is executed in BIC. Window size of CUBIC is increased gradually.

B. Result of Android TCP Tuning

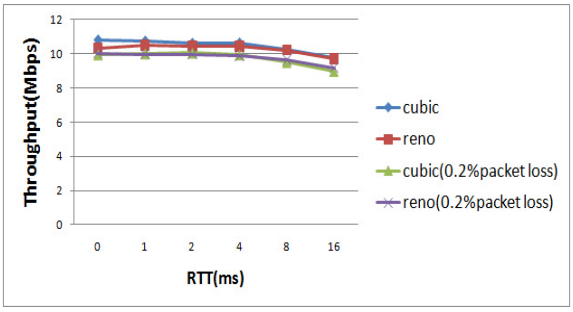


Figure 8. TCP Throughput between Server and Android Terminal with Various Algorithms

Figure 8 shows the result of performance that each Congestion Window control algorithms are applied to Android.

 ISBN:978-1-61208-113-7

Packets are sent from Android terminal to server. The performance of a round-trip including 0.2% packet loss nearly equal to that of a round-trip with no packet loss. The performance of the case including packet loss has a little decline, although there are no substantial difference between them.

C. Performance of Android to Android Communication

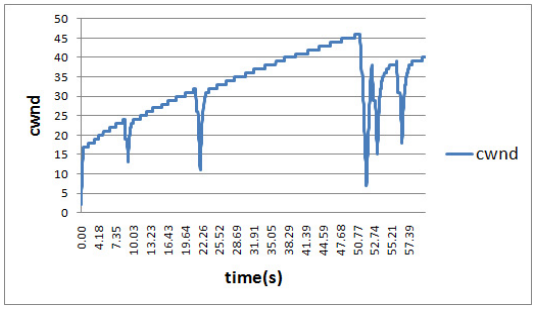


Figure 9. Cubic in RTT=0 [ms] including 0.2% packet loss.

CWND of Android to Android communication is shown in Figure 9 with basic communication by Kernel Monitor. Figure 9 shows that CWND degrades sometimes during the communication. Since the receiver side is Android terminal in this experiment, this is considered to be resource shortage in some cases.

D. Comparison of CWND size in TCP Tuning

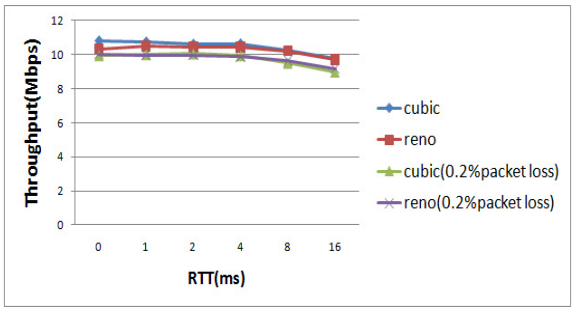


Figure 10. TCP Throughput between Server and Android Terminal with Various Algorithms

Next, we have observed the difference of CWND behavior in the case of TCP Tuning with Kernel Monitor. Figure 10 shows the behavior of CWND with Cubic in RTT=0[ms] including 0.2% packet loss. Figure 11 shows CWND with Reno.

According to Figure 10 and 11, we can observe the way of increasing CWND size is different depending on the Congestion Window control algorithms. Comparing both figures, Reno sets CWND size at half of ssthresh when congestion occurs. Next time when congestion occurs, Reno

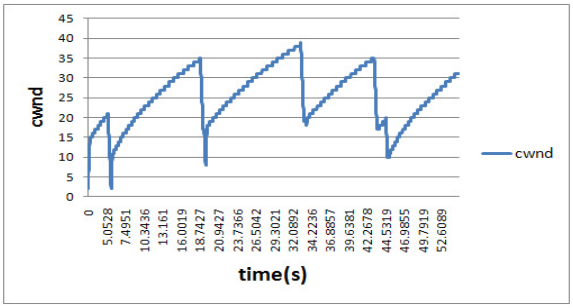


Figure 11. Reno in RTT=0 [ms] including 0.2% packet loss.

restarts CWND size at ssthresh and enters to congestion avoidance phase. Thus, Reno reduces CWND size less than that of Cubic when CWND size is decreased. The way of increasing CWND size is different from another one, although maximum value of CWND is between 40 and 45 in both cases. Throughputs of these algorithms are almost equal. In the case including packet loss, the behavior of CWND can be observed differently from another algorithm.

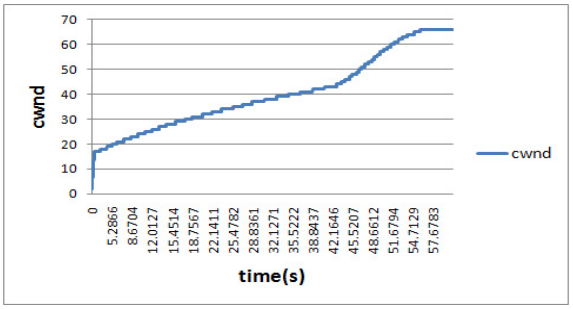


Figure 12. CWND value with Cubic in RTT=0 [ms]

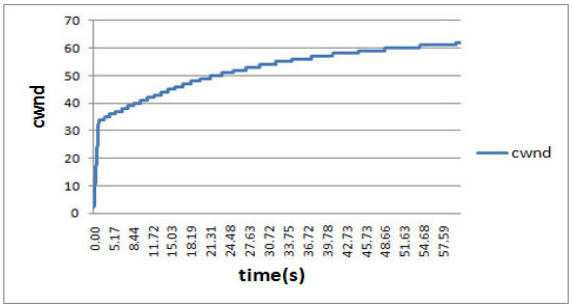


Figure 13. CWND value with Reno in RTT=0 [ms]

Next, the behavior of CWND in both cases of TCP Tuning is compared without packet loss. Although there were no substantial difference between two algorithms without packet loss in our previous work monitoring on Android-x86 [6], Figure 12 and 13 show that the way of increasing CWND is different from each other. This is because Android

mobile phone terminal increase CWND slower than the case of Android-x86.

VI. Conclusions

In this paper, we have developed Kernel Monitor that works on Android terminals. As a result, it becomes possible that kernel of Android can be analyzed with almost the same way as that of the Kernel Monitor of general-purpose PC. In addition, with Kernel Monitor, we have analyzed the relation of throughput and CWND behavior of Android during the communication.

As a future work, we are going to analyze various parameters of Android with Kernel Monitor, find peculiarity of Android, and analyze that behavior in detail. Especially, we are going to investigate into the behavior of simultaneous communication by multiple Android terminals.

Moreover, since most of Android communication is performed by applications on Dalvik, we are going to analyze the specific feature of Dalvik.

References

[1] Android:http://www.google.co.jp/mobile/android

[2] Takashi Majima, Tetsuo Yokoyama, Gang Zeng, Takeshi Kamiyama, Hiroyuki Tomiyama, and Hiroaki Takeda, ”CPU Load Analysis Using Dalvik Bytecode on Android,” IPSJ SIG Technical Reports, March 2010

[3] Reika Higa, Kosuke Matsubara, Takao Okamawari, Saneyasu Yamaguchi, and Masato Oguchi, ”Analytical System Tools for iSCSI Remote Storage Access and Performance Improvement by Optimization with the Tools, ”In the 3rd IEEE International Symposium on Advanced Networks and Telecommunication Systems (ANTS2009), December 2009.

[4] Iperf:http://downloads.sourceforge.net/project/iperf/iperf/2.0.4

[5] Sourcery G++ Lite 2008q-3-72 for ARM GNU/Linux:http://www. codesourcery.com/, http://www.codesourcery.com/sgpp/lite/arm /portal/release644

[6] Miki Kaori, Masato Oguchi, Saneyasu Yamaguchi, ”A study about behavior of Transport layer on Android terminals in a wireless LAN,” In Summer United Workshops on Parallel, Distributed and Cooperative Processing (SWoPP2010),August 2010.

[7] Kojima Akihisa, Ishihara Susumu, ”A Congestion Control of Cooperate with Internetmediate Node for MANET” Multimedia, Distributed, Cooperative, and Mobile Symposium (DICOMO2007),July 2007.

[8] Hasegawa Go, Murata Masayuki, ”Transport-layer protocols for high-speed and log-delay networks” The Institute of Electronics, Enformation and Communication Engineers, Technical Committee Conferences,February 2007.

[9] Sangtae Ha, Injong Rhee, and Lisong Xu, ”CUBIC: A New TCP-Friendly High-Speed TCP Variant” ACM SIGOPS Operating Systems Review, Volume 42 Issue 5, pp.64-74, July 2008.

[10] Habibullah Jamal and Kiran Sultan, ”Performance Analysis of TCP Congestion Control Algorithms” International Journal of Computers and Communications, Issue 1, Volume 2, pp.30-38, 2008.

 ISBN:978-1-61208-113-7

Path Selection in WiMAX Networks with Mobile Relay Stations

Pavel Mach, Zdenek Becvar
Department of Telecommunication Engineering, Faculty of Electrical Engineering
Czech Technical University in Prague
Technicka 2, 166 27 Prague, Czech Republic
machp2@fel.cvut.cz, zdenek.becvar@fel.cvut.cz

***Abstract*—Introduction of mobile relays into networks based on IEEE 802.16 standard brings new challenges. The paper proposes signaling mechanism for acquisition of channel state information for mobile relays and provides detail analysis of the amount of signaling overhead caused by their introduction. In addition, the investigation whether the connection of the mobile users through the mobile relays enhances the system performance is carried out. The obtained simulation results indicate that by means of mobile relays, the overall throughput can be increased and signaling overhead reduced.**

Keywords-CSI, mobile relays stations, path selection, WiMAX.

I. INTRODUCTION

Over the several last years, wireless systems and technologies established themselves as one of the fastest growing and developing area in the field of telecommunications. Especially IEEE 802.16 standards, also known as WiMAX (Worldwide Interoperability for Microwave Access), have a great potential. In 2004, IEEE 802.16-2004 [1] version intended for fixed users was approved, which was followed by IEEE 802.16e [2] finished one year later. To cope with increasing users' requirements for higher data rates, new WiMAX working groups were established in 2006, i.e., IEEE 802.16j [3] and IEEE 802.16m [4]. The IEEE 802.16j version introduces Relay Stations (RS), which enhances system capacity and increases network coverage. The main aim of IEEE 802.16m is to improve spectral efficiency and to minimize signaling overhead. According to [5], three types of RSs are defined; fixed, nomadic and mobile RSs. The fixed RS (FRS) is permanently installed at the same location. Although the nomadic RS (NRS) is also fixed when operating, its position can be changed as needed. The last type of the RS, i.e., the mobile RS (MRS) is moving in similar way as Mobile Stations (MS).

When RSs are introduced into WiMAX based networks, several routes between the MS and Base Station (BS) can be found. The challenge is to select the route offering the best network's performance. Data can be routed either through implementation of certain existing routing protocol or by modification of signaling at MAC layer. Nonetheless, the implementation of whole routing protocol into WiMAX seems to be too complicated. Thus, the preferable solution is to make simple modifications at MAC layer instead. In the scope of IEEE 802.16j standardization body, several proposals focus on routing issues in relay-based WiMAX networks (e.g., in [6][7]). In [6], the authors propose a signaling mechanism for efficient routing intended for IEEE 802.16j standard. Nevertheless, the best point of attachment is decided immediately after the network entry procedure and no potential changes during MS's operation are discussed. In [7], end to end routing and connection management is addressed. Besides the standardization activity, a lot of research papers dealt with routing issues in IEEE 802.16 networks with the FRSs. The common aim is to design effective path selection metrics and to propose suitable path selection algorithms for appropriate routing of data (see, e.g., [8]-[13]).

To our best knowledge, all existing works dealing with routing issues assume only the FRSs, not the MRSs. Thus, the novelty of this paper is that takes into consideration also the MRSs and analyzes their impact on system performance. The first objective is to propose signaling mechanism allowing acquisition of channel state information (CSI). If the CSI of individual routes is known, the most appropriate path between the MS and BS can be selected. As a basis, the proposal uses the signaling mechanism introduced in [11], which takes into consideration only the FRSs, and extends it for the MRSs as well. The second objective is to investigate if it is profitable to use the MRS by MSs that are in close vicinity of the MRS but not located on the same vehicle as MRS.

The rest of the paper is organized as follows. The next Section overviews MAC management messages used in the proposed signaling scheme and contemplates the assumptions considered in this paper. A brief description of principle explained in [11] is introduced in Section 3 together with proposed signaling mechanism taken into consideration MRSs. In addition, the analysis of overhead introduced by signaling mechanism is addressed. Section 4 and section 5 describe simulation scenario and simulation results respectively. The last Section gives our conclusion.

 ISBN:978-1-61208-113-7

II. PRELIMINARIES

A. MAC management messages

In IEEE 802.16e standard, several options to obtain CSI between the BS and MS are defined. In our proposal, the BS acquires the CSI by means of MOB_SCN-REP (mobility scanning report). The MOB_SCN-REP contains the results of scanning procedure. Time allocated for the scanning and reporting period is allocated to the MS through MOB_SCN-RSP messages (mobility scanning response). Two types of reporting are specified: a) event triggered reporting and b) periodic reporting. In the event triggered reporting, the MS sends the reports after each measurement of channel parameters, i.e., CINR (Carrier to Interference and Noise Ratio), RSSI (Received Signal Strength Indicator), Relative delay and RTD (Round Trip Delay). In the periodic reporting, the reports are sent periodically.

If the MS is attached to the BS through one or more RSs, the results of scanning have to be retransmitted to the BS. One option is to simply send the MOB_SCN-REP received by the individual MSs. Nevertheless, this option generates significant amount of signaling overhead. The second option is to combine obtained scanning results by all subordinate stations into one message labeled as MOB_RSSCN-REP. The definition and structure of the MOB_RSSCN-REP can be found in [14].

Similarly as in [11], the proposal distinguishes the activity/inactivity of the MS. To be more specific, the scanning and reporting periods depend on whether the MS has data to send or not. When the MS becomes active, the bandwidth request header (in IEEE 802.16 standard labeled as BW request) is send to the BS.

B. Assumptions

The IEEE 802.16 standard specifies several physical layers. In the paper, physical layer based on Orthogonal Frequency Division Multiple Access (OFDMA) is considered. In addition, the uplink and downlink transmissions utilize the same frequency band, i.e., the Time Division Duplex (TDD) is assumed. The maximal number of hops between the MS and BS is restricted to three hops. Consequently, the MRS can be attached either directly to BS or via one FRS.

According to [5], the MRS is supposed to be placed on some kind of public traffic vehicle such as bus or tram. Hence, the MRSs can be considered as another MSs moving along the predefined trajectory (e.g., the path, along which the bus is traveling from departure to terminal station). The only difference with regard to the MS is that the MRS generates distinguishable more traffic as aggregates traffic of its subordinate MSs. In comparison with the MS, the MRS is assumed to be active all the time since its control information at the beginning of every frame must be transmitted. Consequently, no active/inactive state is distinguished as in case of MS. In addition, the MSs located at the same vehicle as the MRS are supposed to be fixed with respect to this particular MRS.

III. PROPOSED SIGNALING METHOD

This section firstly describes the signaling mechanism, which purpose is to obtain CSI. Secondly, the impact of MRSs in the network on the signaling overhead is analyzed. Thirdly, the path options for MSs are contemplated.

A. CSI acquisition with Mobile Relay Stations

When the MS is attached to the FRS or MRS while in inactive state, the acquisition of CSI is done exactly in the same way as described in [11]. Hence, the scanning period is set to t_1 and reporting period is set to t_2 (see Fig. 1). Both the scanning and reporting periods (scheduled in the MOB_SCN-RSP), are derived from the speed of the MS and can occur relatively infrequently in order to save valuable radio resources. In addition, to further minimize signaling overhead, the value of t_2 can be set to $n*t_1$ where n is the integer value. Nevertheless, the MOB_SCN-REP should be sent by the MS anytime if CINR between the MS and the access station (BS or RS) drops below a specific value for a certain amount of time. This principle guarantees that a handover can be made in advance. If the MSs are connected to the MRS, the MRS creates single MOB_RSSCN-REP message by combining of all MOB_SCN-REP messages and retransmits it in the direction of BS. Thus, the reduction of signaling overhead is achieved. Nevertheless, if the MS is attached to the FRS, the FRS itself simple relays the message toward the BS. This is due to the fact that the reporting intervals of individual users attached to the FRS are not necessarily scheduled at the same time intervals and to wait for all MOB_RSSCN-REP can result in outdating of reporting information.

On the other hand, the MRS itself needs to send CSI to the BS in order to use the optimum route to the BS. As the MRS is considered to be active all the time, the value of scanning period t_3 and reporting period t_4 should be set to

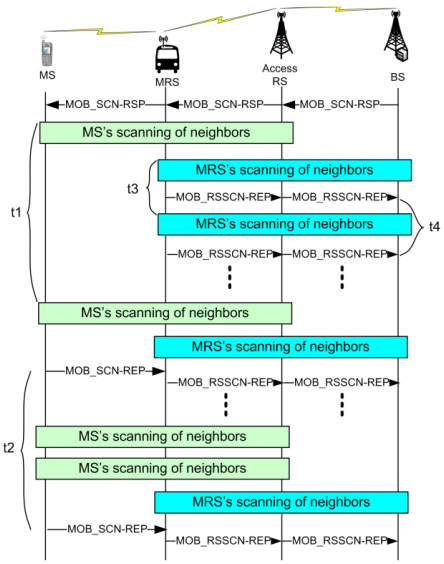


Figure 1. Scanning and reporting periods of MRS and MS (MS in inactive state).

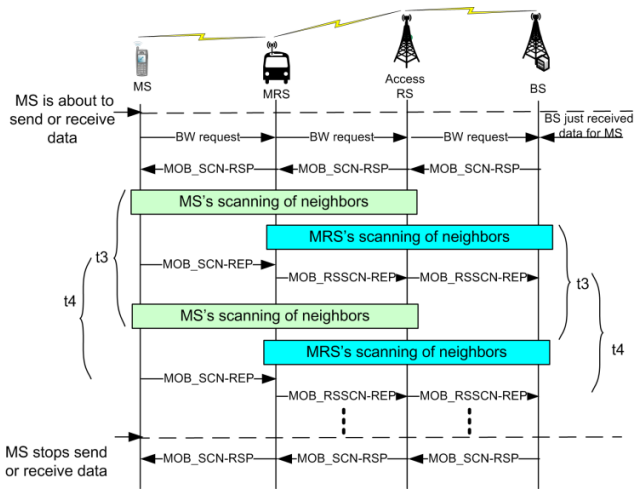


Figure 2. Scanning and reporting periods of MRS and MS (MS in active state).

much shorter values than t_1 and t_2. Thus, the up to date route between the MRS and BS can be maintained. The results acquired during the MRS's scanning interval are appended to MOB_RSSCN-REP message.

If MS becomes active, the scanning and reporting intervals should be changed accordingly in order to obtain up to date channel information (see Fig. 2). The BS learns about MS's transition from the inactive to the active state either through BW request, which originates at the side of the MS, or when the BS has some data designated to this MS (see Fig. 2). Consequently, the BS transmits another MOB_SCN-RSP message to schedule new scanning and reporting intervals. The MS performs scanning at t_3 interval while the reporting interval is set to t_4 similarly as in case of permanently active MRS. However, if the MS is located at the same vehicle as the MRS, the reporting and scanning intervals for the MS can remain the same as described in Fig. 1. The reason is that the MS is fixed (or slowly moving) with regard to the MRS. Thus, the variation of channel conditions is supposed to be minimal and signaling overhead can be reduced.

B. *Analyzis of signaling overhead*

The amount of overhead introduced by the acquisition of CSI between the BS and MSs is proportional to several parameters. Note that by signaling overhead is meant the overhead introduced by proposed mechanism, not the whole overhead at MAC layer. The first parameter corresponds to the size of the reporting messages MOB_SCN-REP (ms_1) and MOB_RSSCN-REP (ms_2). In general, the number of MS's neighbors has direct impact on the messages size. Since the MOB_SCN-RSP message is sent infrequently, its impact on the overhead is minimal and it is neglected in the paper. The second parameter is the amount of active MSs in the system (n). The third parameter is the number of hops between MS_i and the BS (noh_i), i.e., how many times the reporting messages have to be relayed to reach the BS. The last parameter influencing the overhead is the system configuration's setting (e.g., reporting period rp_i, nominal channel bandwidth, OFDMA parameters, etc.).

The overhead is composed of several parts. The first part of the overhead is caused by MOB_SCN-REP message sent by the MSs to their access stations, which could be expressed as:

$$OH_{MS}\left[\frac{b}{s}\right] = \sum_{i=1}^{n_1} ms_1^i \times \frac{1}{rp_i(t_2)} \qquad (1)$$

where n_1 is the number of inactive users and n_2 represents the amount of active users (i.e., $n_1 + n_2 = n$). Note that message size ms_1 is expressed in bits and reporting periods in seconds. The second part of the overhead is generated by the RSs. For the MSs connected only through the FRSs, the overhead can be formulated as follows:

$$OH_{FRS}\left[\frac{b}{s}\right] = \sum_{i=1}^{n_{11}} ms_1^i \times \left(noh_i^{MS} - 1\right) \times \frac{1}{rp_i(t_2)} + \sum_{i=1}^{n_{21}} ms_1^i \times \left(noh_i^{MS} - 1\right) \times \frac{1}{rp_i(t_4)} \qquad (2)$$

where n_{11} is the number of inactive users attached to the FRSs, n_{21} represents the amount of active users connected to the FRSs and noh_i^{MS} is the number of hops between the MS i and BS. In other words, the MOB_SCN-REP message is simple relayed to the BS as described earlier. The overhead caused by the MSs connected to the MRS and MRSs itself can be expressed as:

$$OH_{MRS}\left[\frac{b}{s}\right] = \sum_{i=1}^{m} ms_2^i \times noh_i^{MRS} \times \frac{1}{rp_i(t_4)} \qquad (3)$$

where m is the number of MRS and noh_i^{MRS} corresponds to the amount of number of hops between the MRS i and BS. The size of MOB_RSSCN-REP varies depending on the amount of received MOB_SCN-REP sent by subordinate MSs, which could be formulated as:

$$ms_2[b] = K + (n_{22} + 1) \times (ms_1 - K) + K_1 \qquad (4)$$

where K is the size of message fields that are transmitter disregarding the amount of received MOB_SCN-REP messages (n_{22}), K_1 stands for the information added by the RS in order to recognize by the BS, which MSs are sending reporting information. In [14], it is demonstrated that if at least two messages are combined at the side of MRS, saving of overhead is achieved.

 ISBN:978-1-61208-113-7

C. *Path selection options*

The MS can be connected either directly to the BS, to the FRS or to the MRS. The question is whether the MS situated near of the MRS (but not at the same vehicle as the MRS) can use the MRS to access the BS as indicated in Fig. 3. On one hand, the overall system throughput may be enhanced since the route via the MRS could offer better connection to the users. On the other hand, the connection through the MRS may have drawback since the route between the MS and BS can change rapidly, e.g., the advantage of attachment through the MRS is only of temporary duration. In this regard, higher number of MS's handover initialization may occur. Nonetheless, the excessive number of handovers can be mitigated by utilization of HDT (Handover Delay Timer) technique proposed in [15] which purpose is to delay handover initialization.

The purpose of following simulation is to investigate whether the connection through the MRSs can enhance overall system performance, which is measured in terms of system throughput and the amount of signaling overhead.

IV. SIMULATION SCENARIO

The simulations are done in MATLAB environment. The parameters' setting is given in Tab. 1. The simulation model is composed of one BS and eight FRSs. A deployment of individual stations is illustrated in Fig. 4. In the simulation, four MRS moving along predefined rectangular trajectories are considered (the initial position of MRSs is also shown in Fig. 4)

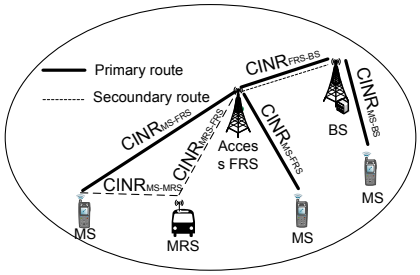


Figure 3. Path selection option for the MS.

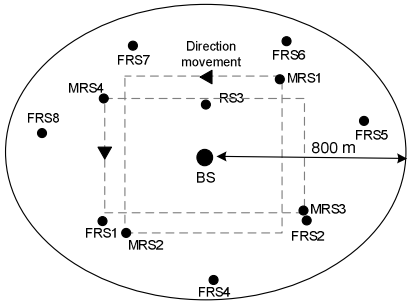


Figure 4. Deployment of RSs and MRS within BS cell.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Frequency band [GHz]	3.5
Channel bandwidth [MHz]	20
Number of MS	50
MS and MRS velocity [m/s]	10-50
Frame duration [ms]	10
BS transmit power P_t [dBm]/height [m]	43/30
FRS transmit power P_t [dBm]/height [m]	30/30
MRS transmit power P_t [dBm]/height [m]	30/4
MS transmit power P_t [dBm]/height [m]	23/2
Noise [dBm]	-100.97
Simulation time [min]	60

The MRS can be connected either directly to the BS or via one intermediate FRS. Connection through more than one FRS is not allowed due to maximum number of hops restrictions (which is up to 3 hops between the MS and BS). Additionally, attachment of MRS through another MRS is not considered.

The movement of the MSs is managed as follows. At the beginning of simulation, an initial position of each MS is randomly determined so that the MS has to be located within defined range, i.e., between 0 to 800 m from the BS. Additionally, random movement direction is determined for all individual MSs in the system. The mobile terminal is moving along straight line until the distance from the BS is equal or larger than defined BS's cell area. In such circumstance, a new direction of the MS is established. This mechanism guarantees that no MS roams out of the BS range during the simulation process.

Two path loss models taken from [16] are implemented. The first one is suitable for LOS communication and describes radio channel behavior between the BS-FRS and the FRS-FRS. The second one is assigned for NLOS communication between the BS-MS, FRS-MS, BS-MRS, FRS-MRS and the MRS-MS.

The path between the MS and BS is selected according to the minimum Radio Resource Cost (RRC) metric (more detail may be found in [17]).

The reporting period t_4, assumed in the simulation, corresponds to the optimal reporting period for active MS/MRS derived from [11]. If the MS is inactive, the reporting period t_2 is set to value of $t_4/10$ in order to minimize signaling overhead.

The system performance is analyzed in terms of system throughput and overhead generated by the MSs and MRSs. To that end, three scenarios are considered. The first scenario represents the situation when MRSs are not assumed (in the following figures labeled as "Scenario A"). Nonetheless, some of the MSs' are positioned at public traffic vehicle moving along predefined trajectories. In the second scenario, the MRSs are installed at public traffic vehicle (in the following figures labeled as "Scenario B"). Thus, the MS situated at the bus are connected to the network right through newly deployed MRS. In the last scenario, it is assumed that also MSs currently not placed at

 ISBN:978-1-61208-113-7

TABLE II. HDT SETTING FOR SCENARIO C

Scenario type	HDT value [s]
Scenario C1	0.01
Scenario C2	0.1
Scenario C3	0.5
Scenario C4	1
Scenario C5	5

TABLE III. TRAFFIC MODELS TYPE

Model type	**VoIP**	**FTP**	**HTTP**
VoIP only	100%	0%	0%
Traffic Mix I	30%	30%	40%
Traffic Mix II	10%	80%	10%

the MRS can use this MRS as an access station (in the following figures labeled as "Scenario C"). Scenario C considers different values of HDT as shown in Tab. 2.

To evaluate the maximal throughput, a full queue traffic model is implemented [18]. The throughput evaluated in the paper represents a system WiMAX capacity obtained at the MAC level. Hence, the overhead introduced by higher layer protocols (e.g., network, transport, etc.) is not considered.

To estimate the amount of the signaling overhead due to reporting, the size of MOB_SCN-REP and MOB_RSSCN-REP messages are derived from [2] and [14] respectively. The activity and inactivity of MSs depend on implemented traffic models. In the simulation, VoIP only and two traffic mixes are considered as indicated in Tab. 3 (detail traffic models description can be found in [18]).

V. SIMULATION RESULTS

Fig. 5 shows the normalized signaling overhead caused by the reporting messages MOB_SCN-REP and MOB_RSSCN-REP depending on the MSs/MRSs velocity. The worst performance is obtained by Scenario A as the highest amount of overhead is generated for all traffic models. The difference in the amount of generated signaling overhead for individual traffic models is caused by diverse ratio of MSs' activity/inactivity. Thus, in case of VoIP model, the MSs are in inactive state much more often than in case of Traffic Mix I/II (i.e., reporting period is more often set to t_2 instead of t_4). The Fig. 5 further demonstrates that by introduction of MRSs into the system, the size of reporting overhead can be reduced (see Scenario B in the Fig. 5). The maximal achieved reduction is obtained for Traffic Mix II, which is approximately 30% when compared to Scenario A. As already explained, the minimization of the overhead is possible due to two facts: i) the MS located at the same moving vehicle as MRS can set its reporting period to t_2 independently on the activity/inactivity and ii) the MRS is able to combine received MOB_SCN-REP messages into one message. Although, the amount of signaling overhead is increased by utilization of Scenario C, the results are still better than in case of Scenario A (especially if Traffic model II is used). The reason for the increase of signaling overhead (when compared to Scenario B) is that the MSs connected through MRS are usually connected to the BS via more hops.

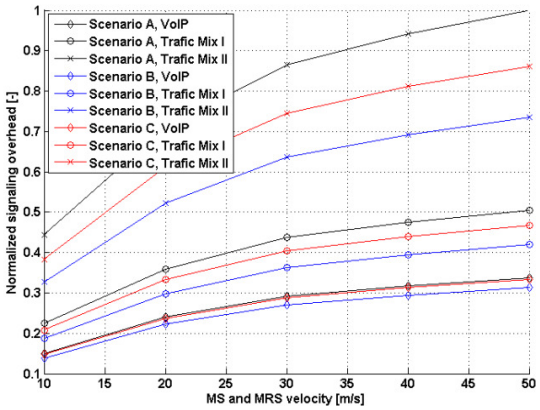


Figure 5. Signaling overhead due to reporting.

Fig. 6 illustrates how many handovers are performed per simulation run depending on MSs/MRSs velocity. The best results are achieved by Scenario B as the overall number of handovers is reduced by utilization of MRS (approximately by 50% for MSs/MRSs' velocity of 10 m/s and by 34% for MSs/MRSs' velocity of 50 m/s). The less number of handovers in comparison to Scenario A is acquired due to the fact that the MSs currently positioned at the MRS do not perform handovers. If the Scenario C is implemented, distinguishable increase of initiated handovers is observed. The reason is that in some cases the attachment via the MRS is only temporaly. Nevertheless, this drawback can be mitigated by implementation of HDT. When HDT value is set to 5 s, the overhead generated by executed handovers is comparable to scenario A.

Fig. 7 presents the throughput achieved for all investigated scenarios depending on offered traffic load. In case of Scenario A, already at middle traffic load, not all data could be transmitted to the destination station. The better results are achieved for Scenario B (improvement by 9.2 %) and for Scenario C1 (improvement by 14 %).

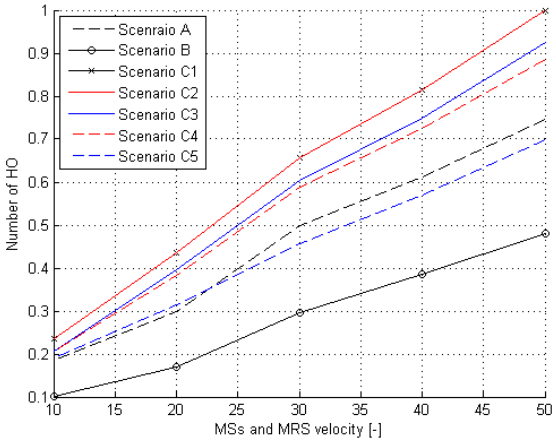


Figure 6. Number of MSs' handover per simulation run.

 ISBN:978-1-61208-113-7

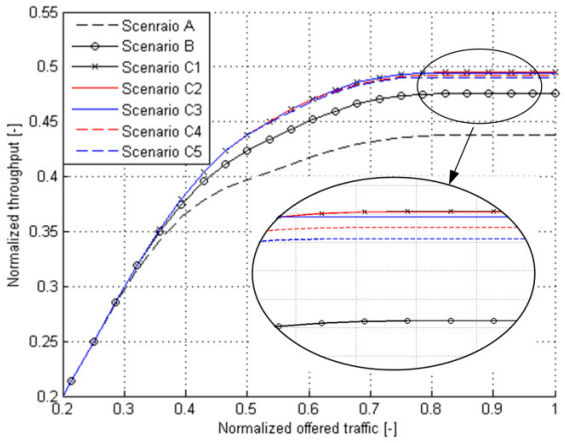


Figure 7. Normalized system throughput in dependence on offered traffic.

Decrease of system throughput by application of HDT is only marginal. From Fig. 6 and Fig. 7 can be also derived the optimal HDT value, which is 5 s as the number of performed handovers are noticeably mitigated while the system throughput is still nearly the same as case of Scenario C1.

VI. CONCLUSION

The paper proposed a mechanism for acquisition of CSI if MRSs are implemented into the network. In addition, detail analysis is done to estimate the amount of overhead generated by reporting messages.

The obtained simulation results indicate that system throughput can be improved since MRSs provide to its MSs better signal quality. In some cases, also the MSs not currently positioned at MRS may utilize connection offered by near MRS. In this way, the system throughput can be further enhanced. Nevertheless, this option has some drawbacks, i.e., number of performed handovers can be significantly increased as the connection through the MRS is only temporally. In order to overcome this issue, HDT is implemented. By utilization of HDT, the excessive number of HO is decreased while the system throughput is still nearly unaffected. The results also demonstrated that the signaling overhead generated by reporting of scanning information can be reduced by means of the MRS.

The disadvantage of MRSs' introduction can be seen in potential increase of interference and longer packet delays if the MSs, not currently located at the same moving vehicle as the MRS, connect to it as described in the paper. Thus, in future work we would like to addresses these issues.

ACKNOWLEDGMENT

This work has been performed in the framework of the FP7 project ROCKET IST-215282 STP, which is funded by the EC. The Authors would like to acknowledge the contributions of their colleagues from ROCKET Consortium (http://www.ict-rocket.eu).

REFERENCES

[1] IEEE Std 802.16-2004, IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.

[2] IEEE Std 802.16e-2005, IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, 2006.

[3] IEEE 802.16j, Baseline Document for Draft Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, 2007.

[4] IEEE802.16m Task Group: IEEE 802.16m proposals. Resources documents. http://www.ieee802.org/16/tgm/, 2009.

[5] J. Sydir, IEEE 802.16 Broadband Wireless Access Working Group – Harmonized Contribution on 802.16j (Mobile Multihop Relay) Usage Models, July 2006.

[6] K. Baum, M. Cudak, S. Ramachadran, P. Satori, and E. Visotsky, "Signaling for Efficient Routing," paper no. C802.16j-06_159r1, available online at <http://ieee802.org/16>, 2007.

[7] G. Q. Wang at al., "MMR Network end-to-end routing and connection management," paper. no. C802.16j-07/092, available online at <http://ieee802.org/16>, 2007.

[8] D. M. Shrestha, S.-H. Lee, S.-Ch. Kim, and Y.-B. Ko, "New Approaches for Relay Selection in IEEE 802.16 Mobile Multi-hop Relay Networks," Lecture Notes in Computer Scienece, pp. 950-959, 2007.

[9] S.-S. Wang, Y.-H. Tsai, H.-Ch. Yin, and S.-T. Sheu, "An Effective Path Selection Metric for IEEE 802.16-based Multi-hop Relay Networks," 12th IEEE Symposium on Computers and Communications (ISCC), pp. 1051-1-56, 2007.

[10] S. Ann, K. G. Lee, and H. S. Kim, "A Path Selection Method in IEEE 802.16j Mobile Multi-hop Relay Networks", International Conference on Sensor Technologies and Applications (SENSOECOMM), pp. 808-812, 2008.

[11] P. Mach, Z. Becvar, and R. Bestak, "Acquistion of Channel State Information for Routing Purposes in Relay-based WiMAX Networks," 8th International Conference on Networks (ICN), p. 170-175, 2009.

[12] D. Li and H. Jin, "Relay selection in two-hop IEEE 802.16 mobile multi-hop relya networks," 1st Internation Workshop on Education Technology and Computer Science (ETCS), pp. 1007-1011, 2009.

[13] S. Ann and H. S. Kim, "Relay association method for optimal path in IEEE 802.16j mobile multi-hop relay networks," *European Transactions on Telecommunications*, 2010.

[14] Z. Becvar and P. Mach, "Reduction of Scanning Reporting Overhead in IEEE 802.16 Networks with Relays," In 9th International Conference on Networking (ICN), pp. 109-114, 2010.

[15] Z. Becvar and J. Zelenka, "Implementation of Handover Delay Timer into WiMAX," 6th Conference on Telecommunications (CONFTELE), pp. 401-404, 2007.

[16] IEEE 802.16j, Multi-hop Relay System Evaluation Methodology (Channel Model and Performance Metric) paper No. 06/013r3, 2007.

[17] Z. Becvar, P. Mach, and R. Bestak., "Initialization of Handover Procedure in WiMAX Networks," ICT-MobileSummit, 2009.

[18] IEEE 802.16m, Evaluation Methodology Document, IEEE 802.16m paper No. 08/004r2, 2008.

 ISBN:978-1-61208-113-7

Planning with Joint Clustering in Multi-hop Wireless Mesh and Sensor Networks

Y. Drabu and H. Peyravi
Department of Computer Science
Kent State University
Kent, Ohio 44242
email: ydrabu@cs.kent.edu and peyravi@cs.kent.edu

***Abstract*—Wide spread deployment of wireless mesh networks for broadband access requires careful deployment and planning in terms of laying down the network infrastructure. Deploying such networks comes with some major inter-related issues including capacity planning, scalability and access reliability. Planning includes determining the number of gateways, optimal placement of gateways and relay nodes, maximizing coverage while minimizing the operational cost.**

This paper focuses on a planning approach that aims at increasing access fairness and fault tolerance using an overlapping clustering technique. It provides alternate paths for nodes residing at the edge of the clusters and mitigates upstream blocking towards the gateways to control delay, congestion and loss rate.

***Keywords*-wireless mesh network; clustering; wireless deployment; gateway placement;**

I. Introduction

The demand for seamless broad-band wireless Internet access has been the major driving force behind the development of multi-hop wireless mesh networks (WMNs) [1]. Wireless mesh networks combine several existing technologies and concepts from cellular, ad hoc, and sensor networks to improve network coverage, easy of deployment and better throughput. The shared wireless nature of the medium makes them more susceptible to failure. Transmission link failure, in which a wireless link experience an excessive loss rate or prolonged delays, is a common case of failure. This is mainly due to outdoor noise, interferences, multi-path fading, contention and congestion. Wireless mesh networks may also be subject to a variety of other faults including faults in network elements and protocol faults [2]. These faults result in low throughput and excessive delay or no connection at all. Some faults that are supposed to recover from path failure may create routing loops or a black holes. However, these faults can be mitigated or prevented by a careful and robust planning.

For a successful deployment of WMNs in such an environment, it is essential to provide certain resilience to the network connectivity during planning and deployment to avoid potential failures [3]. WMNs planning involves several interdependent factors that include network topology, network coverage, traffic demand, and capacity assignment. The optimal number of gateways and their locations have to be determined in advance and before deployment.

Gateway placement has a significant impact on the overall network performance including its financial viability and access reliability. In WMNs, traffic congestion is mostly due to up-stream aggregate traffic heading towards a gateway and that can be controlled by proper placement of gateways. While minimizing the number of gateways will reduce the deployment cost, fewer gateways will increase the average hop distance and consequently increases the average delay and the average relay load of the intermediate routers. Finding the optimal number of gateways can formulate as an optimization problem in which an objective function minimizes the number of gateways subject to a set of QoS requirements.

The gateway placement problem is similar to the clustering problem. Clustering has been studied extensively in the context of operation research with different objective functions and optimization goals. One of the main distinctions among clustering techniques is in their objective function. In the context of wireless mesh works, a set of more complicated and dynamic objective functions is involved in the clustering The set could include cluster size, number of clusters, hop count, and relay load. Given a significant portion of delay a packet suffers is associated with the hop count the packet travels, it is important to put a limit on the hop count towards a gateway and then optimize the other objectives.

In this paper, the optimal layout of the network has been integrated to the planning phase of a WMN deployment. It will allow diverse routing and fault-tolerant provisioning, particularly for links that face higher blocking probability to access a gateway. Generally, an edge node of a cluster faces more blocking probability and hence higher delay and loss rate than nodes closer to the cluster head. A new clustering technique is introduced by which the gateway placement algorithm allows redundant cluster membership to improve access reliability while keeping the optimality intact.

The rest of the paper is organized as follows. Section II summarizes related work with respect to the gateway placement problem in wireless mesh networks. Section III covers basic preliminaries and definitions. Section IV presents a new network clustering techniques based on maximal independent sets. Section V extends the clustering algorithm of Section IV for joint cluster membership for disadvantaged nodes. Section VII concludes the paper.

 ISBN:978-1-61208-113-7

II. Related Work

Link failure in a wireless network is commonly caused due to interference in the medium or traffic congestion and on rarer occasion due to the radio malfunction.

Fault tolerance has been getting a lot of attention in the area of sensory networks [2], due to the higher node failure rate, large scale of the network and the desire to increase automation. Fault recovery in such networks has been addressed in terms of routing [4], topology control [5], power assignment [6] and channel assignment [7], [8].

On the other hand, fault tolerance in wireless mesh networks, which are more stable than sensory networks, has been studied in the context of networking layer using routing protocols [9]. The routing protocols finds an alternate path to route a packet from a source to a destination if the primary path fails. However, all routing algorithms assume some route redundancy in the underlying network topology, which is more apparent in WMNs than in sensor networks.

In [10] the authors discuss fault tolerance with respect to gateway placements. To address node and link failures they modify the gateway placement LP formulation and add a fault tolerance constraint to ensure over-provisioning via multiple independent paths. They propose a greedy heuristic to address gateway placement that iteratively picks up nodes that increasingly satisfy the traffic demand without necessarily selecting a node that satisfies the most demand. Therefore, in this work we focus on building wireless network that are fault tolerant at the network topological level.

A. Gateway Placement Problem

In the following, we give an overview of the gateway placement problem and provide the most common approaches proposed in the context of wireless mesh networks.

1) Placement with Integer Linear Programming: The optimal placement can be obtained by minimizing the number of clusters(k in Equation 8) subject to a set of constraints such maximum cluster size, maximum cluster radius, etc. The combinatoric algorithm checks all possible combinations to find a solution that satisfies all the QoS constraints. This approach is prohibitively expensive and does not scale beyond very small network.

2) Placement with Greedy Approach: In [10], the authors suggest to place the gateway simply at a location where it satisfies the most traffic demand subject to the capacity of the gateway and relay nodes. However their greedy approach can lead to an imbalance loading of certain gateway and does not support all the quality of service requirements.

3) Placement with Iterative Clustering: The earliest work that directly addressed the placement of gateways in a wireless mesh network [11] describes the problem as a capacitate facility location problem with additional constraints. The author solves the placement problem by breaking it into two sub problems. First a polynomial time approximation algorithm that cuts the network into disjoint clusters using a shifting algorithm or a greedy dominating independent set algorithm. Once the initial clustering is completed, each cluster is evaluated to that ensure QoS constraints are met. If the QoS is violated, then the cluster is sub-divided into smaller clusters at the node where QoS is violated. However, for the solution to work, it is assumed that the underlying medium access protocol is TDMA(Time-Division Multiple Access). TDMA protocols require synchronization, which is hard to achieve in large multi-hop wireless networks. Additionally, the proposed solution generates higher fragmented clusters.

4) Placement with Recursive Clustering: Similar to [11], in [12], the authors form a cluster and a spanning tree within each cluster to obtain a near optimal solution. They propose a recursive algorithm that builds a clustering and then admits it into the solution only if it meets the QoS constraints. The algorithms is able to produce lesser number of clusters than those in [11]. However, the cluster sizes have a large variance and the clustering does not vary uniformly when with a uniform change in QoS constraints.

5) Split-Merge-Shift: The Split-Merge-Shift [13] starts with an initial clustering graph and then it goes through a few iterations of *Split*, *Merge*, and *Shift* operations to form the final clustering. The algorithm does not necessarily generate an optimal solution initially, but over a set of iterative Split-Merge-Shift operation it converges close to optimality.

III. Preliminaries

In planning, deployment or updating a wireless network, it is often necessary to determine the transmission range with an acceptable throughput. While there are many factors that affect the tranmission range, the theoretical transmission distance can be obtained from a few key specifications.

Definition 1 (Transmission Range): Given the transmission power P_t, the receiving power P_r, the transmission range d can be calculated as,

$$d = \frac{\lambda}{4\pi}\sqrt{\frac{P_t G_t G_r}{P_r F_t}} \tag{1}$$

where G_t, G_r are the transmitting and receiving gains with an acceptable loss factor F_t and λ is the wavelength of the communication channel.

Definition 2 (Transmittance Matrix): We define the binary transmittance matrix $T = [t_{ij}]$ as

$$t_{ij} = \begin{cases} 1 & \text{if} \quad d_{ij} \leq t_r, \quad i \neq j \\ 0 & \text{otherwise.} \end{cases} \qquad 1 \leq i, j \leq N \tag{2}$$

where d_{ij} be the Euclidian distance between node i and node j obtained from Equation 1.

Definition 3 (Reachability Matrix): The h-hop binary reachability matrix $R_h = [r_{ij}]$ is defined as

$$R_h = T^1 \vee T^2 \vee \cdots \vee T^h = \bigvee_{k=1}^{h} T^k, \tag{3}$$

where $\vee$ is the binary OR operation, and

$$r_{ij} = \begin{cases} 1 & \text{if node } i \text{ is at most } h \text{ hops away from} \\ & \text{node } j,\ i \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Definition 4 (Hop Count Matrix): The entries of the hop count matrix $H = [h_{ij}]$ give the hop distance between nodes within the reachability range such that,

$$h_{ij} = \begin{cases} k & \text{if node } j \text{ is within } k \leq h \text{ hops from node } i \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

where $h = \max\{h_{ij} \mid 1 \leq i, j \leq N\}$.

Corollary 1: The reachability matrix $R_h = [r_{ij}]$ can be obtained from the hop-count matrix H as,

$$r_{ij} = \begin{cases} 1 & \text{if } h_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Definition 5 (Cluster): A cluster $C(V', E') \subseteq G(V, E)$ is an acyclic sub graph of G such that, $V' \subseteq V$ and $E' \subseteq E$.

Definition 6 (Clustering): A clustering is a way of partitioning graph $G(V, E)$ and can be formally defined as a set of clusters Ω, where,

$$\Omega = \{C_1, C_2, \cdots, C_k\}, \quad 1 \leq k \leq N, \tag{7}$$

with the following properties:

$$\begin{cases} P_1: & \bigcap_{i=1}^{k} C_i = \emptyset \\ P_2: & \bigcup_{i=1}^{k} C_i = G \\ P_3: & V(\Omega) = V(G) \\ P_4: & E(\Omega) = \bigcup_{i=1}^{k} E(C_i) \subseteq E(G). \end{cases} \tag{8}$$

Property P_1 guarantees that clusters are independent with no nodes in common. Relaxing this property allows overlapping clusters. Ω can be represented by an $N \times N$ asymmetric binary matrix with k non-zero rows, each representing a cluster with exactly a 1 on each column, characterizing each node to belong only to one cluster. Formally,

$$\Omega_{N \times N} = [\omega_{ij}] \in \{0, 1\}$$

with the following constraints,

$$\begin{aligned} (a): & \quad \sum_{i=1}^{N} w_{ij} = 1, \quad 1 \leq j \leq N \\ (b): & \quad \sum_{j=1}^{N} w_{ij} > 0, \quad \text{for some } i \end{aligned} \tag{9}$$

Constraint (a) guarantees that each node belongs only to one cluster, and constraint (b) makes node i as a cluster head with its member nodes j, where $\omega_{ij} = 1$ $(1 \leq j \leq N)$.

Later, in Section V, we relax the property P_1 in Equation 8 and its corresponding constraint (a) in Equation 9 to allow a node to participate in more than one cluster.

Generally, clustering formation and optimal placement of gateways (cluster heads) with some QoS constraints is known to be an $\mathcal{NP}$-hard problem [14]. Several heuristic are proposed in [11], [10], [12] and [13] to place gateways efficiently in a given network. However limited work has been done to allow fault-tolerant through joint memberships.

While ad hoc routing algorithms such as AODV (Ad hoc On-demand Distance Vector) and some of its variations can be used to route packets in multi-hop wireless mesh networks, generally they face a few shortcomings when directly applied to WMNs. First, their throughput performance does not typically scale to meet the expectation, particularly for real-time applications that are delay-sensitive or even loss-sensitive for data transmission. Their effective performance in terms of QoS requirements such delay, loss and jitter depends strongly on the underlying topology and the transmission range. Second, unlike ad hoc networks, where the traffic flows between arbitrary nodes, WMN traffic is either to or from a designated gateway (similar to a cellular system). A WMN routing algorithm must exploit this property to gain efficiency, which is the intention of this paper. Third, ad hoc routing algorithms are designed to deal with the possibility of highly mobile nodes and that requires a significant amount of overhead for route discovery, mobility and maintenance. On the other hand, WMNs routers have minimal mobility. This is yet another characteristic that can be exploited for efficiency. Finally, in terms of planning, ad hoc network planning is mostly done manually without any systematic approach, and often without paying attention to the overall system cost.

Because of their relatively fix position (or change of position is limited within a certain range) of WMN nodes, the implication is that the routing paths can be created that are likely to be stable. This will substantially reduce the routing overhead. The most commonly used topology for WMNs is a grid layout which is due to the layout of building and blocks. The relatively stationary topology of WMNs suggests that we can develop a more simplified routing algorithm along with a systematic approach to the planning and deployment. All these necessitate a different approach to the planning, deployment, and routing in WMNs which is the focus of this paper.

IV. Planning with Disjoint Clustering

By strictly applying property P_1 in Equation 8 along with constraint (a) in Equation 9, a clustering matrix in the form of Equation 2 can be formulated to represent an optimal set non-overlapping clusters covering the mesh network.

One of the important QoS requirements in WMNs is to determine the maximum number of hops a packet can travel before reaching its intended destination (gateway). For that, we form the h-hop reachability matrix R_h from Equation 3 that identifies the reachability set for each node on its rows.

This can be viewed as an initial clustering (trivial clusters) in which every node is considered to be a cluster head with all its members within h-hop distance. Clearly, this will create the maximum possible number of clusters (N) with maximum overlap amongst them. However, condition P_1 in Equation 8 is not satisfied for non-overlapping clusters. To satisfy property P_1, we introduce a *cluster graph* in which cluster C_i is connected to cluster C_j if $C_i \cap C_j = \emptyset$, $1 \leq i, j \leq N, i \neq j$. We further define the corresponding *clustering overlap matrix* as follows.

Definition 7 (Clustering Overlap Matrix): The entries of the clustering overlap Matrix, $O = [o_{ij}]$ is defines as,

$$o_{ij} = \begin{cases} \sum_{k=1}^{N} r_{ik} \wedge r_{jk} & i \neq j \\ 0 & i = j \end{cases} \quad 1 \leq i, j \leq N, \tag{10}$$

where $\wedge$ is the binary operation AND, and o_{ij} is the inner product of row i and row j of R_h. In effect o_{ij} gives the number of common nodes in two adjacent clusters headed by nodes i and j are considered two cluster heads. We define the adjacency clustering matrix $A = [a_{ij}]$ that describes relationships between clusters as follows.

Definition 8 (Clustering Adjacency Matrix): The clustering adjacency matrix $A = [a_{ij}]$ is defined as,

$$a_{ij} = \begin{cases} 1 & \text{if } o_{ij} \geq I_c \neq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

For disjoint clustering, first, we consider the case where $I_c = 1$. We start with the transmission matrix T in Equation 2 and a maximum clustering radius of h. We compute the reachability matrix within h-hop distance for each node according to Equation 3. The clustering adjacency matrix A identifies the relationship between potential clusters in terms of node sharing. We define matrix A' as the complement of A where,

$$a'_{ij} = \begin{cases} 1 & \text{if } a_{i,j} = 0 \\ 0 & \text{if } a_{i,j} = 1 \end{cases} \quad (12)$$

A' identifies all *pair-wise* disjoint clusters.

Definition 9 (Inter Cluster Distance): Inter cluster distance D_h is defined as the maximum number of hops between any two clusters.

To find the optimal location of cluster heads with maximum coverage, one has to find the maximum clique (maximum complete subgraph) of the graph associated with the adjacency matric A'. We use the Algorithm original developed by [15] to find the largest clique (complete subgraph). The current implementation of the algorithm searches for maximal independent vertex sets in the complementer graph. Given we have applied the constraint of hop-count h on each cluster, depending on the network topology, the algorithm does not necessarily cover all the nodes in the clustering.

Consider the 100-node mesh network of Figure 1. The initial

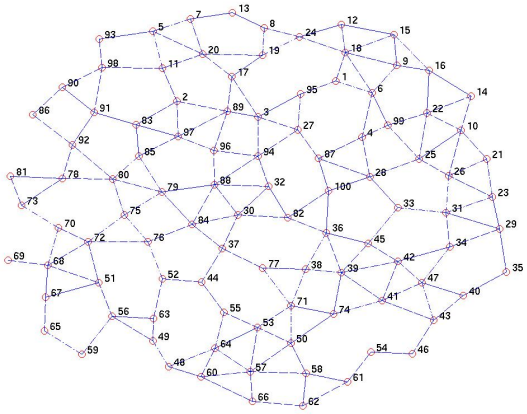


Fig. 1. A 100-node mesh network.

clustering is shown in Figure 2 in which 8 gateways optimally cover the network with maximum hop count $h = 2$. The initial clustering does not cover nodes all the nodes mainly due to the *inter-cluster* constraints applies. For example, nodes $\{41, 45, 48, 52, 55, 57, 60, 64, 66, 73, 79, 81, 82\}$ have not been assigned to any of the clusters due to: (i) the maximum 2-hop coverage ($h = 2$) by the cluster heads, and (ii) the inter-cluster distance $I_c = 1$, i.e., neighboring clusters are at least one hop away from each other. However, uncovered nodes

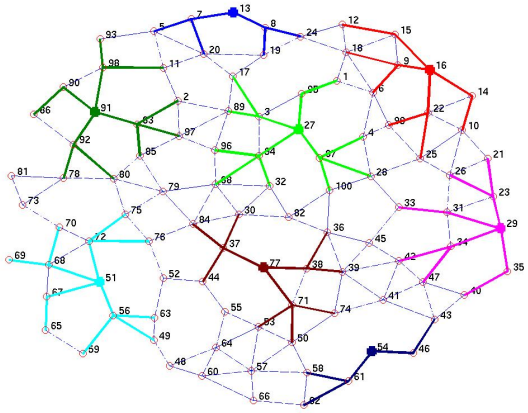


Fig. 2. Initial clustering with $h = 2$ and I_c=1.

nodes are at most h hops away from a nearby cluster. For that we identify the *inter-cluster* distance matrix for the above clustering algorithm.

Theorem 1: A node is either a cluster head or at most $2h$ hops away from a cluster head.

Proof: Given, the clustering algorithm forms only disjoint clusters, there are two cases.

If $\bigcup_{i=1}^{k} C_i = G$, then the clustering algorithm covers all nodes in the network and Property P_2 holds. Every node is within h hops from a cluster head and no nodes lies between two adjacent clusters, and hence $D_h = 1$.

If $\bigcup_{i=1}^{k} C_i \neq G$, then there is at least one node that does not belong to any of the clusters. Let $v \in G$ but $v \notin \bigcup_{i=1}^{k} C_i$ be such a node. Let the closest cluster to v be C_i with its cluster head node u. The h-hop reachability set of v is either disjoint or it has some nodes in common with the h-hop reachability set of u. Let $R_h(v)$ and $R_h(u)$ be the h-hop reachability sets for node v and u, respectively.

- Case 1: $R_h(v) \cap R_h(u) \neq \emptyset$. Let w be a common node in both reachability sets. Then the hop distance $H(v, w) \leq h$ and the hop distance $H(u, w) \leq h$. Hence $H(u, v) \leq 2h$.
- Case 2: $R_h(v) \cap R_h(u) = \emptyset$. Then v by itself constitutes an independent reachability set within its h radius and forms an independent cluster.

■

From Theorem 1, we can conclude the following corollaries.

Corollary 2: The maximum inter-cluster distance $D_h = h$.

Corollary 3: A node that has not been assigned to any clusters is at most h hops away from a neighboring cluster.

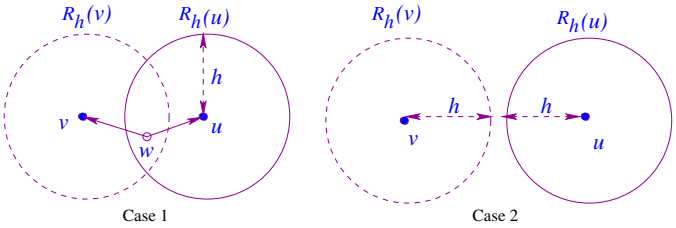


Fig. 3. Inter-cluster distance.

This is shown in Figure 2 in which nodes $\{41, 45, 48, 52, 55, 57, 60, 64, 66, 73, 79, 81, 82\}$ are either one hop or two hops away from a neighboring cluster, where $h = 2$. After the initial clustering, we will find the nearest cluster for the remaining nodes them to join. This is shown in Figure 4. Therefore, the radius of the final clustering is at

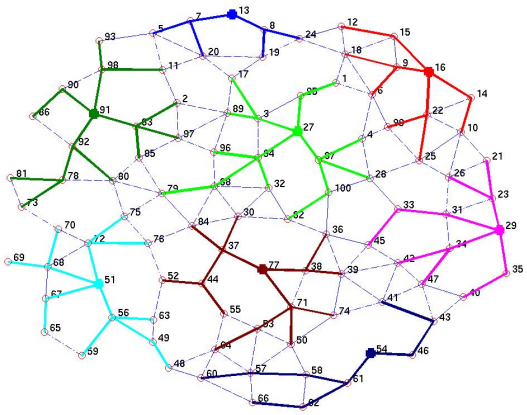


Fig. 4. Final disjoint clustering, $2 \leq h \leq 4, I_c = 1$.

most $2h = 4$ hops. While the inter-cluster distance $I_c = 1$ is one hop among the neighboring clusters, due to the network topology some clusters are affected by the residual nodes left out from the constraint h in algorithm 1.

Algorithm 1: Disjoint Clustering

Input : Transmittance Matrix T, h, $I_c = 1$
Output: Array C of cluster heads

1 Calculate $R_h = \bigvee_{k=1}^{h} T^k$ Eqn. 6
2 Calculate $o_{ij} = \sum_{k=1}^{N} r_{ik} \wedge r_{jk}$ Eqn. 10
$1 \leq i, j \leq N$
3 Calculate A from O for $I_c = 1$ Eqn. 11
4 Calculate A' from A
5 Use the maximal independent set [15] to identify the cluster heads.
6 Form clusters by incorporating the reachability set (from R_h) for each cluster head.
7 Assign nodes outsides clusters to the closest cluster.

A. Analysis

The processing time involved in Steps 1-4 in Algorithm 1 are all based on two-dimensional matrices (mostly sparse matrices) and bounded by $O(N^2)$. The processing time and memory space in step 5 are bounded by $O(N + m)$ and $O(Nm\delta)$, respectively, where N is the number of nodes, m is the number of edges and δ is the maximal independent sets of the graph [15].

V. Planning with Joint Clustering

By relaxing property P_1 in Equation 8 and constraint (a) in Equation 9, we can obtain clusters that can share available bandwidth at the edge of clusters. Nodes at the edge of clusters belong to more than on cluster simply because they are in disadvantage positions as far as gateway access is concerned. They can dynamically switch their cluster membership due to a weak or bad connection at the edge of each cluster. This can be achieved in two ways; i) making $D_h = 1$ and allow the inter-cluster links be shared by the neighboring clusters, or ii) make clusters overlap by one or more hops. Note that the objective of this paper is to compensate access disparity with access redundancy for those nodes further away from a gateway to improve their throughput.

In this clustering scheme, nodes that are h hops away from a gateway have memberships in more than one cluster. The joint clustering algorithm is simply an extention of disjoint clustering algorithm with inter-cluster nodes having at least dual membership in neighboring clusters. This is shown in Algorithm 2. The difference between Algorithms 1 and 2 are

Algorithm 2: Joint Clustering

Input : Transmittance Matrix T, h, $I_c \geq 1$
Output: Array C of cluster heads

1 Calculate $R_h = \bigvee_{k=1}^{h} T^k$ Eqn. 6
2 Calculate $o_{ij} = \sum_{k=1}^{N} r_{ik} \wedge r_{jk}$ Eqn. 10
$1 \leq i, j \leq N$
3 Calculate A from O for a given I_c Eqn. 11
4 Calculate A' from A
5 Use the maximal independent set [15] to identify the cluster heads.
6 Form clusters by incorporating the reachability set (from R_h) for each cluster head.
7 Assign nodes outsides clusters to adjacent clusters.

in steps 3 and 7. Figures 5 and 6 show one hop ($h = 1$) clustering with $I_c = 1$ and $I_c = 2$, respectively. Similarly, Figure 7 for $h = 2$ and $I_c = 3$ The choice for h and I_c depends on the planning. Clearly, increasing I_c reduces the number of clusters and hence the number of gateways and higher fault-tolerance. The drawback is the amount of delay.

VI. Performance

In multi-hop networks, the throughput performance of a connection decays exponentially with an increase in hop count.

 ISBN:978-1-61208-113-7

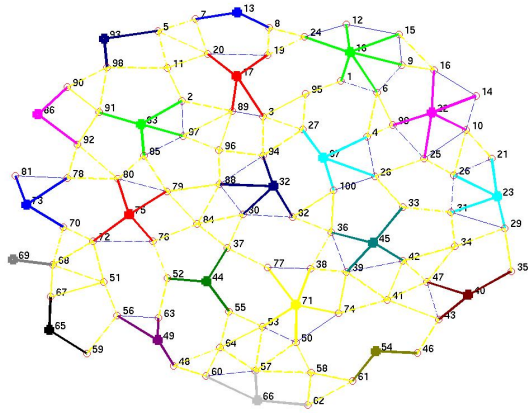


Fig. 5. $h = 1, I_c = 1$

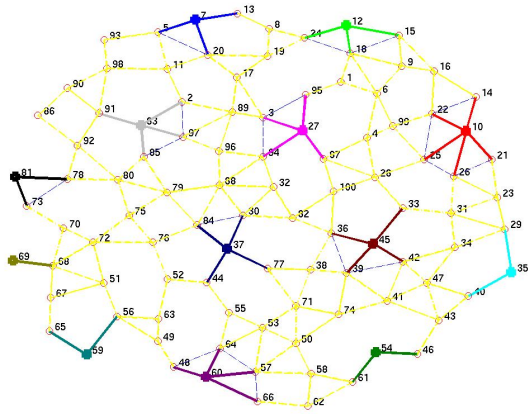


Fig. 6. $h = 1, I_c = 2$

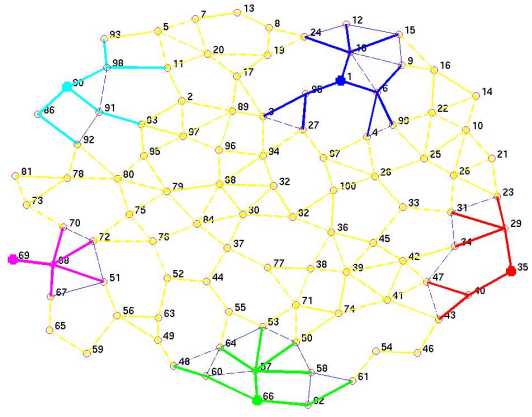


Fig. 7. $h = 2, I_c = 3$.

This is illustrated in Figure 8 for a simple network that is illustrated in Figure 9 in which a packet hops towards a gateway. A packet may runs into successive contentions and that results in higher blocking probability on each hop along the path towards its intended gateway. Each link carries a local traffic load (ρ) and relays up-stream traffic from previous nodes.

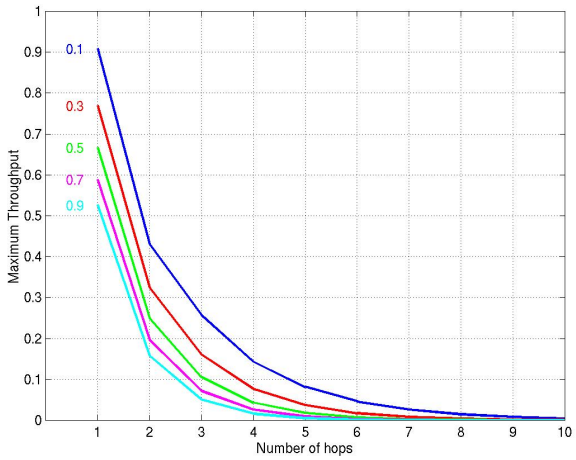


Fig. 8. Maximum throughput performance across different traffic loads.

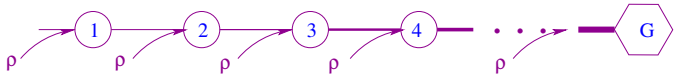


Fig. 9. A linear multi-hop network.

Figure 8 illustrates the theoretical end-to-end throughput as a function Erlang blocking probability for different traffic load (ρ) excluding loss rates. While Erlang blocking probability has been studied extensively in the content of switching networks and telephony, it can be used to approximate the blocking probably for applications such as VoIP in packet switching networks or wireless cellular systems, in which the end-to-end is connection-oriented. The exponential throughput degradation has also been observed in several experiments we conducted with Roofnet [16] which is discussed in Section VI-A, and simulation results we obtained in Section VI-B.

A. *Roofnet Experiment*

In our Roofnet experiment, a 5-node mesh network was created in a 3D indoor environment. The configuration of wireless network is depicted in Figure 10. With the Roofnet protocol [16] installed on each router (54 GHz, 802.11g), we measured the effect of the number of hops as well as

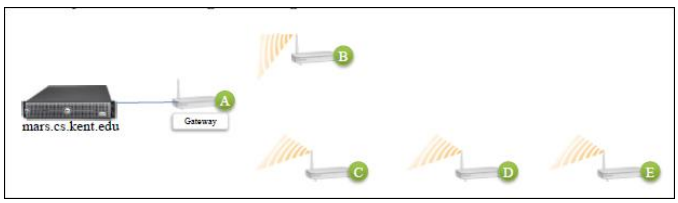


Fig. 10. A small Roofnet configuration.

the hop distance (in dB) between each pair of relay routers. File with various sizes destined towards the gateway (node A in Figure (10) were generated by the clients that are connected to a neighboring relay router. Each experiment was conducted five times at different times of the day and the results were averaged. The experiment was then repeated by increasing the hop distance. We also varied the Euclidian distance within each hop. Figure 11 illustrates the exponential decay of throughput performance when a hop distance (left) or the hop count (right) increases. The two major observation

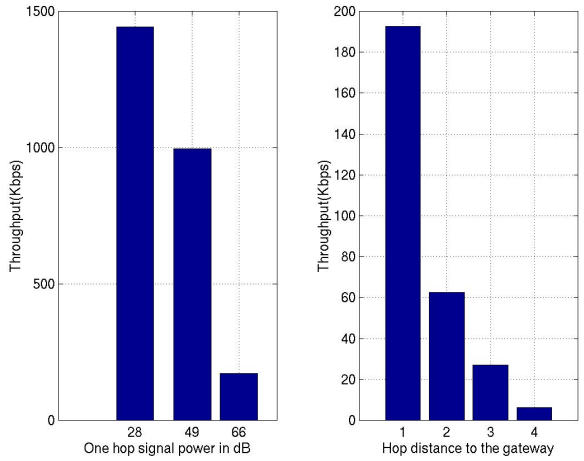


Fig. 11. The effect of hop distance (left) and hop count (right) on throughput.

from Figure 11 are: (i) the transmission range between routers has to be factored into the clustering and deployment, as incorporated in Algorithm 1, and (ii) the throughput disparity for nodes distant away from the gateway has to be mitigated, as incorporated in Algorithm 2.

B. Simulation Experiment (Qualnet)

In addition to Roofnet experiment, we also used Qualnet [17] simulator to perform two separate experiments. The first was to study the effect of hop distance on throughput and access fairness, and the second was to see the effect of providing alternate cluster membership for a source node on the edge of a cluster as proposed in this paper to increase network resilience with respect to failure.

In the first experiment, we setup five wireless nodes as a linear multi-hop network, similar to the network in Figure 9, with the first node being the traffic source and the gateway node being the traffic destination. We then increased the load, by increasing the traffic generated on the source node and observed the effect of load on the throughput. We repeated the simulation with traffic flows between source destination pairs. In this experiment, we limit the hop-counts to 4, as the throughput performance deteriorates significantly beyond 4 hops. Figure 12 shows the throughput performance varying based on the load for different hop distances. We observed that the throughput of the flows with fewer hop count is significantly better than the throughput of flows with higher hop count. As load increases beyond 40-50%, the throughput for all scenarios start decreasing, mainly due to the contention resolution and back-off algorithms provisioned in the 802.11 protocol. This result is in line with our analysis and the Roofnet experiments.

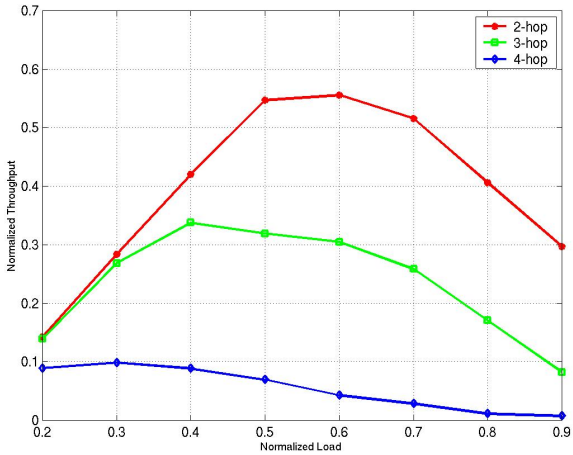


Fig. 12. Load vs. throughput with hop distance.

In the second experiment as shown in Figure 13, we allowed overlapping of clusters in the planning phase, thus enabling node S to be part of cluster C1 and C2. We created a traffic flow from node S to node D. We then studied the effect of load on the throughput with and without the overlapping clustering. Figure 14 illustrates the effect of load on the throughput with

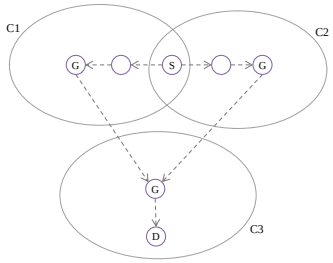


Fig. 13. Overlapping clusters simulation setup.

a node belongs to one or two clusters. The throughput for a node is significantly higher if it belong to two clusters.

VII. Conclusion

Mesh networks, due to their multi-point to multi-point architecture, inherently lend themselves to being more resilient to faults. However, the placement of wired gateways in these WMNs has a significant impact the on network throughput performance, cost and capacity to satisfying the quality of service (QoS) requirements as well as fault tolerance. In the context of gateway placement, the QoS is influenced by the number of gateways, the number of nodes served by each gateway, the location of the gateways, and the relay load on each wireless router.

In this paper we developed a new clustering technique that improves fault-tolerance in wireless mesh networks. It

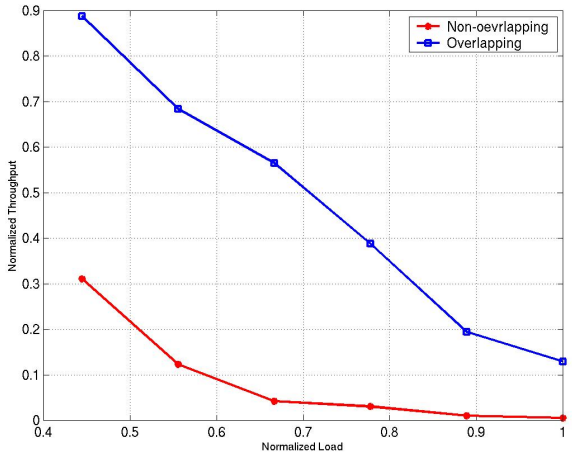


Fig. 14. Load vs. throughput with and without overlapping

also mitigates the throughput disparity among nodes distant way from a gateway by allowing them to join multi-cluster. Simulation results and measurements have shown a significant improvement in terms of throughput once clustering is incorporated during the deployment process. The clustering is independent of underlying network routing protocol, but improves the overall performance.

REFERENCES

[1] P. Bahl, "Opportunities and challenges of community mesh networking," May 2004.

[2] Q. H. Lilia Paradis, "A survey of fault management in wireless sensor networks," Journal of Network and Systems Management, vol. 15, no. 2, pp. 171–190, June 2007.

[3] W. W. Ian F. Akyildiz, Xudong Wang, "Wireless mesh networks: a survey," Computer Networks, vol. 47, no. 5, pp. 445–487, March 2005.

[4] E. Royer and C. Toh, "A review of current routing protocols for ad-hoc mobile wireless networks," Apr 1999.

[5] X.-Y. Li, P.-J. Wan, Y. Wang, and C.-W. Yi, "Fault tolerant deployment and topology control in wireless ad hoc networks: Research articles," Wirel. Commun. Mob. Comput., vol. 4, no. 1, pp. 109–125, 2004.

[6] S. A. et al., "Distributed power control in ad hoc wireless networks," PIMRC, 2001.

[7] W. Si, S. Selvakennedy, and A. Y. Zomaya, "An overview of channel assignment methods for multi-radio multi-channel wireless mesh networks," J. Parallel Distrib. Comput., vol. 70, no. 5, pp. 505–524, 2010.

[8] K. L. E. Law and A. Kohn, "Topology designs with controlled interference for multi-radio wireless mesh networks," in Mobility '08: Proceedings of the International Conference on Mobile Technology, Applications, and Systems. New York, NY, USA: ACM, 2008, pp. 1–6.

[9] R. Draves, J. Padhye, and B. Zill, "Routing in multi-radio, multi-hop wireless mesh networks," in MobiCom '04: Proceedings of the 10th annual international conference on Mobile computing and networking. New York, NY, USA: ACM Press, 2004, pp. 114–128.

[10] R. Chandra, L. Qiu, K. Jain, and M. Mahdian, "Optimizing the placement of integration points in multi-hop wireless networks," Proceedings of IEEE ICNP, 2004.

[11] Y. Bejerano, "Efficient integration of multihop wireless and wired networks with QoS constraints," IEEE/ACM Trans. Netw., vol. 12, no. 6, pp. 1064–1078, 2004.

[12] B. Aoun, R. Boutaba, Y. Iraqi, and G. Kenward, "Gateway placement optimization in wireless mesh networks with QoS constraints," IEEE Journal on Selected Areas in Communications, vol. 24, no. 11, Nov 2006.

[13] Y. Drabu and H. Peyravi, "Gateway placement in wireless mesh networks with qos constraints," Seventh International Conference on Networking (ICN 2008), Nov 2007.

[14] D. B. Shmoys, E. Tardos, and K. Aardal, "Approximation algorithms for facility location problems," in STOC '97: Proceedings of the twenty-ninth annual ACM symposium on Theory of computing. New York, NY, USA: ACM Press, 1997, pp. 265–274.

[15] S. Tsukiyama, M. Ide, H. Ariyoshi, and I. Shirakawa, "A new algorithm for generating all the maximal independent sets." SIAM J. Comput., vol. 6, no. 3, pp. 505–517, 1977.

[16] B. A. Chambers, "The grid Roofnet: a rooftop ad hoc wireless network," MIT Masters Thesis, 2002.

[17] Q. N. Simulator, "http://www.scalable-networks.com/."

 ISBN:978-1-61208-113-7

Anomaly Detection Framework for Tracing Problems in Radio Networks

Jussi Turkka, Tapani Ristaniemi
Department of Mathematical Information Technology
P.O.BOX 35, Jyväskylä University
Jyväskylä, FI-40014, Finland
jussi.turkka@jyu.fi, tapani.ristaniemi@jyu.fi

Gil David, Amir Averbuch
School of Computer Science
Tel-Aviv University
Tel-Aviv, 69978, Israel
amir@math.tau.ac.il, gil.david@yale.edu

Abstract—**This paper shows a novel concept of using diffusion maps for dimensionality reduction when tracing problems in 3G radio networks. The main goal of the study is to identify abnormally behaving base station from a large set of data and find out reasons why the identified base stations behave differently. The paper describes an algorithm consisting of pre-processing, detection and analysis phases which were applied for RRC (Radio Resource Control) connection data gathered from the live radio networks. The results show that the proposed approach of using dimensionality reduction and anomaly detection techniques can be used to detect irregularly behaving base stations from a large set of data in a more self-organized manner.**

Keywords - radio network optimizatio; data mining; anomaly detection; self-organizing networks.

I. INTRODUCTION

Mobile phone service providers need to gather excessive amounts of information from the network to be able to optimize the network performance and solve several kinds of problems in the network operation. Data gathering and analysis can be a rather resource consuming task requiring a lot of manual work, specialized equipment and expertise. As the rapid evolution of the cellular networks and the increased capacity demands has led to a situation where the operators need to maintain several multi-vendor radio access networks (RAN) simultaneously, the burden of operating and maintaining such a complex network infrastructure has caused a need to develop more automated solutions for network deployment, operation and optimization.

Self-organizing networks (SON) and Minimization of Drive Tests (MDT) solutions are widely regarded as prominent approaches which would reduce the operational expenditures and at the same time improve the perceived end-user quality-of-service (QoS). Both approaches are also actively researched by the biggest network vendors and operators in the 3rd Generation Partnership Project (3GPP) making possible the solutions for network deployment automation.

The target of the MDT work item in 3GPP is to define a set of measurements and measurement reporting procedures which would help operators to gather more data from the network without excessive manual drive tests [1]. The MDT study targets to monitor and detect coverage problems in the network such as coverage holes, weak coverage, pilot pollution, overshoot coverage, coverage mapping and UL coverage, to name a few, as described in more details in [2].

On the other hand, the target of the SON work item in 3GPP is to define the necessary measurements, procedures and open interfaces to support self-configuring, self-optimization and self-healing use cases, which can dynamically affect the network operation, and therefore, improve the performance and reduce the manual operation efforts [3]. The SON use cases in [3] targets to coverage and capacity optimization, energy savings optimization, interference reduction, automatic configuration of physical cell identity, mobility robustness optimization, mobility load balancing optimization, random access channel optimization, automatic neighbor relations configuration, and inter-cell interference coordination.

Both SON and MDT use cases describe many measurement quantities. This creates a challenge, which is currently overlooked in the literature: how to effectively post-process the measurement data of *huge volumes*? In 3GPP this is left to be a vendor specific solution. Good introduction to the network monitoring and troubleshooting is found in [4] and comparison of different network monitoring tools is done in [5]. A simple and traditional method to monitor networks is to define a problem, a measurable performance indicator (KPI) and a pre-defined threshold, which indicates whether or not the problem exists [4]. However, there are several challenges if the network performance monitoring is done in this way. Firstly, the complexity of the network infrastructure results in many problems, and therefore, the number of performance indicators increases. This results in large databases and makes the identification of the fault/KPI-associations a much more complex task. Secondly, since the networks are complex and dynamic in nature, it is not straightforward at all to define what is a good threshold or what should be measured and how often. If one has several performance indicators, what is statistically the best one to reveal a certain feature in network behavior? Thirdly, due to the dynamic nature of the networks there is a strong need for predictability of the network behavior, which calls for advanced processing of the network performance data.

In this article we are addressing these challenges by proposing a novel algorithm to analyze a large and complex set of network performance data. The analyzed data is gathered from a live 3G network using trace functionality [6]

 ISBN:978-1-61208-113-7

and it includes connection related measurement data. The subscriber and equipment traces provide very detailed information at call level on specific mobiles allowing advanced monitoring and optimization operations i.e., root cause analysis in troubleshooting, optimization of resource usage and QoS, and radio and core network end-to-end call procedure validation [6]. The proposed algorithm consists of four main phases, which are pre-processing, dimensionality reduction, detection of anomalies and post-processing the results for root cause analysis.

This paper is organized as follow. Section II describes the data, the goal of the data mining procedure and pre-processing of the data in more details. Section III describes the data mining techniques. Section IV summarizes the algorithm and shows the results of the analysis. Finally, sections V and VI discuss the future improvements of the algorithm and draw conclusions.

II. PRE-PROSESSING

A. Motivation for Database Processing

The goal of the study was to find a way to analyze the given data set and find a set of base stations which did not behave similarly compared with the regular behavior of the whole set of the base stations. It is worth of noting that no assumptions were made about the normal behavior. Instead, the measured statistics were used to classify the behavior of the base stations assuming that most of them worked as supposed. Moreover, besides finding the anomalous base stations, the target was to find out reasons, why certain base stations are different. Such an approach would make it possible to create a problem pattern data base which can be used to detect malfunctioning base stations in more self-organized manner. In the beginning, this requires manual efforts and experts knowledge in problem classification. However, when the problem pattern data base exists, it can make the problem detection less time and resource consuming in the long run. This makes the networks more profitable reducing the manual efforts of tracing the problems.

B. Data Description

The database was gathered from a foreign operators 3G network and it consists of call session data. Each sample consisted of 72 features with different kind of data types i.e., call duration, UE type and capability, connection related parameters, last visited sector and site. Samples were collected over the area of several radio network controllers (RNC) and 335 base station sectors. The total number of measured samples was 42215 but the samples were collected over a rather short time period of 8 minutes, which can have an effect to the statistical reliability of the analysis. Eventually, only the cells, which had a large number of connection attempts during the 8 minutes duration, were analyzed, and therefore, some of the cells were not included to the analysis at all.

The pre-processing phase of the dataset consisted of several steps. First, many of the original 72 features were filtered out from the dataset because there were many missing samples and it was unclear how to process the missing features. Second step was to choose a subset of features and convert the data to a form, which can be analyzed with the chosen data mining techniques. The chosen three nominal features were RRC establishment reason, RRC release cause and failure source. Each of the features were categorical consisting of several nominal attributes i.e., RRC establishment reason can be detach, registration and either originating or terminating certain connection type such as a low priority connection as seen in Table I. There were several other categorical features, which were not included to this study such as connection detail, UE capabilities, used release version or RRC success.

RRC success was left out because it plays a heavy role in the traditional problem solving strategy and one of the intentions was to find out whether or not the algorithm can provide similar and comparable results compared with the traditional root cause analysis. The process for the manual problem solving is explained briefly in the following.

First, an engineer filters the data and studies only the cells where the failure ratio exceeds the certain threshold e.g., the cells which have the poorest RRC connection success ratio. Next step is to choose the cell and study only the calls which failed. Based on the failure source, detail, establishment cause and release cause, the engineer gets an idea what might be wrong. Usually at this point, the engineer has knowledge about the cause of the error and they can access in more detailed log files to trace the error source. For the same reason, the above-mentioned features were chosen to be the most interesting for detecting anomalies in this study. In some cases, the mobile type is also a useful feature to trace the root cause of the problems i.e., the software related bugs in mobiles can cause problems and unwanted behavior that is not related to the problems in RAN elements at all.

C. Data Pre-processing

Data pre-processing phase consists of data selection, data filtering, and data transformation to a form which can be analyzed with the data mining algorithms. In the data selection, last cell id and three categorical features were selected for the analysis and rest of the 72 features was filtered out. Because the target was to detect cells which were anomalous and because the difficulties in analyzing the categorical attributes, the data was transformed to a numerical form which can be processed easily and effectively with the data mining algorithms. It is worth of noting that there are ways to analyze mixed data bases consisting of nominal and numerical features as explained in [7]. However, the numerical form of data was preferred in this study. The structure of the filtered database is shown in (1)

$$x_l = [\, a, b, c \,], \quad (1)$$

where the variable x_l is the lth RRC connection attempt recorded during the measurement. the variables *a, b* and *c* are categorical features indicating: the RRC connection *establishment reason* with 15 possible nominal attributes, the

RRC *release cause* with 35 possible nominal attributes and the *failure source* with 10 possible nominal attributes.

In data transmission from categorical data to numerical data, a vector for each feature was created with a length corresponding to the number of possible attributes. Furthermore, the three vectors were merged to one high dimensional vector x_{bs} which length is 60. The x_{bs} is the database sample for base station *bs* showing the count, how many times certain RRC connection establishment, released or failure attribute was measured during the 8 minutes measurement period. The structure of the x_{bs} is shown in (2)

$$x_{bs}=[\, a_1 \ldots a_i,\, b_1 \ldots b_j,\, c_1 \ldots c_k \,], \quad (2)$$

where the variable a_i is the count for how many times the *i*th categorical attribute was present during the measurement period for the categorical feature *a* e.g., RRC connection establishment reason. The variables b_j and c_k are the counts for how many times the *j*th and *k*th categorical features were present for the original categorical features *b* and *c*. The features are listed in Table I.

TABLE I
FEATURE LIST

RRC Establisment Reason	
a1) intr_rat_cell_re_select	b16) radio_link_failure
a2) orig_low_prior_signal	b17) synchronization_failure
a3) registration	b18) srnc_relocation
a4) term_low_prior_signal	b19) orig_background_call
a5) orig_background_call	b20) unspec_failure
a6) orig_high_prior_signal	b21) orig_streaming_call
a7) orig_conversational_call	b22) no_resp_from_rlc
a8) term_high_prior_signal	b23) orig_interactive_call
a9) term_conversational_call	b24) iuv_iu_rel_comm_received
a10) detach	b25) rrc_conn_req_nack
a11) orig_streaming_call	b26) call_re_establishment
a12) orig_interactive_call	b27) rrc_dir_sc_re_est
a13) call_re_establishment	b28) physical_channel_failure
a14) srnc_relocation	b29) no_resp_from_rrc_d
a15) orig_subscribed_tra_call	b30) no_resp_from_iuv
RRC Release Reason	b31) fail_in_r_if_proc
b1) pre_emption_failure	b32) rq_ci_ip_not_supp
b2) orig_low_prior_signal	b33) synchronazion_fail
b3) registration	b34) timer_expired
b4) intr_rat_cell_re_select	b35) orig_subscribed_tra_call
b5) term_low_prior_signal	**RRC Failure Source**
b6) nwk_optimisation	c1) bts
b7) orig_high_prior_signal	c2) default
b8) orig_conversational_call	c3) rnc_internal
b9) no_error	c4) cell_reselection
b10) serv_req_nack_from_rm2	c5) radio_interface
b11) rl_setup_failure	c6) iu
b12) detach	c7) transmissio
b13) term_conversational_call	c8) ms
b14) term_high_prior_signal	c9) frozen_bts
b15) inter_system_hard_ho	c10) ciphering

However, there is one very fundamental issue, which must be taken into account in the data mining and the knowledge pattern analysis. Always when one does the data selection and transformation, the information what the data includes can change and get biased depending on the actions what are done? This must be kept in mind whenever doing the data pre-processing and analyzing the results from reliability and validity point of view.

III. DATA MINING

A. Data Mining Principles

Data mining is a process for extracting interesting, previously unknown and potentially useful information patterns from large datasets [8]. The data mining process consists of several phases being data cleaning; data base integration; task relevant data selection; data mining; and pattern evaluation [8]. Data cleaning, integration and selection are data pre-processing phases where the data is prepared for the data mining [8]. The data mining consists of several functions such as classification of the data; association of data; clustering the data; dimensionality reduction and anomaly detection to mention a few [8]. In pattern evaluation phase, the information patterns are visualized and analyzed to see how usable, novel, valid and reliable the findings are i.e., even though something interesting is found, it doesn't mean that it is already usable or useful. In this study, dimensionality reduction and anomaly detection data mining techniques were used.

B. Dimensionality Reduction with Diffusion Maps

This section describes the diffusion maps framework that was introduced in more details earlier in [7][9][10][11]. Diffusion maps and diffusion distances provide a method for finding meaningful geometrical descriptions in high dimensional data sets consisting of points in R^n where *n* is large e.g., 60 in this study. The diffusion maps construct coordinates that parameterize the dataset and the diffusion distance provides a local preserving metric for this data [10]. The data is parameterized by using a graph *G* and weight function kernel *W* measuring the pairwise similarity of the training set of the points in R^n [10].

If a proper kernel is used, then *W* can be normalized into a Markov transition matrix *P* and the most significant eigenfunctions of the Markov matrices provide a good low dimensional geometric embedding in a way that the ordinary Euclidean distance in the embedding space measures the meaningful diffusion metrics of the data [10]. Moreover, the diffusion distance between two points, as described in [9], reflects the geometry of the dataset as well. The expression of the diffusion distance can be given as in [11]

 ISBN:978-1-61208-113-7

$$D_t^2(x,y) = \sum_{k\geq 0} \lambda_k^{2t} \left(v_k(x) - v_k(y)\right)^2, \quad (3)$$

where k is the number of the most significant eigenvectors and variable λ is the eigenvalue of the kth eigenvector. The variables $v_k(x)$ and $v_k(y)$ are the kth right eigenvectors of transition matrix for points x and y. Euclidean distance between two points in the embedded space R^k represents the distances and similarity of the same points in high dimensional space. If two points are close then the diffusion distance in (3) is small. Furthermore, the diffusion distances and the density can be used to detect anomalous base stations as explained later.

C. Anomaly Detection

In anomaly detecting, samples can be either clustered to several categories as in [7][10] or to normal and abnormal samples. The diffusion distance metric in (3) and the diffusion coordinates can be used for the clustering since it reflects the similarity of the points in the original high dimensional space. If the density of a certain point is large in the embedded space then it has many neighbors nearby and it is considered to be regular. In contrary, irregular points have small density. For each point in the embedded space, a k-ball with certain radius is defined and the density is calculated by using any normalized density function based on the samples which lay inside the ball. Since the scale of the each diffusion coordinate in the embedded space is different, the ball radius is scaled as well. The density d_m of the mth point is defined in (4)

$$d_m = \frac{\eta_m}{\sum_{i=1}^{M} |\eta_i|}, \quad (4)$$

where η_m is the number of points inside the ball and the sum in the denominator is norm-1 over all M points.

In this study, the difference between the normal and the anomalous sample was based on the statistical properties of the density distribution of the dataset. A point was abnormal in case the density was smaller than the threshold shown in (5)

$$\mu_d - 2\sigma_d, \quad (5)$$

where variables μ_d and σ_d are the median value and the standard deviation of the point density distribution.

IV. RESULTS

A. Summary of the Algorithm

The used algorithm consists of the following steps:

1. *Dataset pre-processing*: The dataset feature selection and transformation resulted in an input matrix size of 129x60 showing the count of how many times nominal attributes were present for the selected three categorical features.

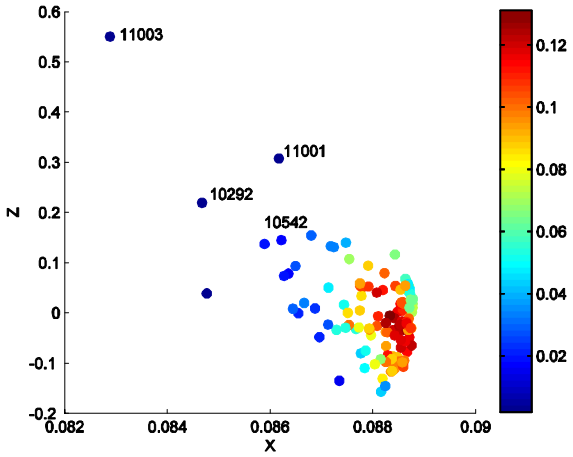


Figure 1. Base station dataset visualization in 3D embedded space.

2. *Dimensionality reduction*: Diffusion map framework was used to reduce the dimensions of the input matrix to size of 129x12 assuming that 12 eigenvectors represents the relevant data for 129 cells. The used pairwise distance metric was L2 and the diffusion epsilon factor 8. The decaying of eigenvalues proof that the findings can be visualized reliably in small dimensional space.
3. *Anomaly Detection*: Difference between normal and abnormal samples was made based on the point densities in the embedded low dimensional space according to (4).
4. *Root Cause Analysis*: An analysis of the reasons why the abnormal base stations are different from the regular ones was made based on the statistical properties of the abnormal samples in high dimensional space after the anomaly detection.

Figure 1 shows the visualization of the data in embedding space using the three most significant eigenvectors. The coloring of the points shows the measured density as described in (4). The dark blue points are irregular samples. There were totally 14 irregular base stations detected. The detection accuracy was rather high compared with the detection strategy descried in Section II which would be based on the observation of RRC connection failure ratio.

For four sectors out of 14, the RRC failure ratio was higher than 5% and this can already trigger an optimization in the network. On the other hand, not all the problems in the network are caused due to the unsuccessful calls, and therefore, it is interesting to study why some of the cells with good RRC success ratio are anomalous. It should be noted that the statistical reliability of the above mentioned observation may require more samples per base station than the calls recorded during the 8 minutes interval.

Moreover, what is the root cause of the problem and how to optimize the network based on the findings? The root cause analysis was done by comparing the 60 features of the irregular base stations one by one to the features of the regular base stations. The features with largest difference

 ISBN:978-1-61208-113-7

TABLE II
ABNORMAL SECTOR PROPERTIES

Cell ID	11003	11001
Failure %	49.4%	43.8%
Calls	2040	937
Feature 1	a) intr_rat_cell_re_select	a) registration
Feature 2	a) registration	a) term_conversational_call
Feature 3	b) pre_emption_failure	b) pre_emption_failure
Feature 4	b) no_error	b) no_error
Feature 5	c) bts	c) bts
Feature 6	c) cell_reselection	c) cell_reselection
Cell ID	10292	10542
Failure %	6.5%	5.9%
Calls	107	153
Feature 1	b) no_resp_from_rlc	b) synchronization_failure
Feature 2	b) physical_channel_failure	b) rrc_conn_req_nack
Feature 3	c) radio_interface	c) frozen_bts
Feature 4	c) ms_c	-

compared with the regular behavior in all cells was used as an indicator to select which features are different i.e., possibly causing the anomalous behavior.

B. Results of Base Station Detection

Table II shows the results of four abnormal sectors with the high RRC failure rates. Sectors 11003 and 11001 had a critically bad RRC success ratios and a high number of call attempts during the 8 minutes period. Both cells seem to be rather congested and there might not be enough resources to create connections. The *pre-emption* property is related to releasing core network resources for creating radio access bearer connections [12]. Therefore, a large amount of pre-emption failures can indicate a poor capacity planning or lack of network resources. However, to find the root cause and a solution to the problem, a deeper analysis of the sector behavior would be needed to verify these assumptions.

Sector 10292 had as well a rather poor RRC success ratio. However, the number of call attempts was small and therefore the total number of failed calls was small. In this sector, the anomalies were due to the number of calls which had events *no_resp_from_rlc* and failures in *radio_interface*. In sector 10542, the anomalies were due to the number of calls which had events *rrc_conn_req_nack* and failures in *frozen_bts*. The details of the other irregular cells are not listed here since the three most abnormal cells already give the idea about the output of the results.

V. CONCLUSION AND FUTURE WORK

In this paper, a novel concept was introduced by using data mining techniques for tracing problems in radio networks. The data mining techniques were used to identify irregularly behaving cells from a large RRC connection dataset. The proposed algorithm consists of pre-processing; dimensionality reduction; anomaly detection; and root cause analysis, which were used with the real 3G network data. The algorithm detected base stations with high failure rate.

For future work, there are ways to improve the accuracy of the algorithm even more. Firstly, the data selection can be extended to include more features. The size of the input matrix affects to the measurement interval, and therefore, the usability of 8 minutes measurement period is to be clarified. Moreover, in this study, the anomaly detection was done in base station domain. On the other hand, the anomaly detection can be done in time domain as well. However, observing the behavior of a single base station over longer time period requires longer measurements. If longer measurements can be conducted, then the analysis can be extended to an automatic discrimination between good and bad base stations in a similar way the discrimination is done in [13]. Furthermore, the root cause analysis can be improved as well. In this paper a simple approach was chosen to indicate what high dimensional features are causing the irregularities. However, an efficient way to choose the meaningful features from the data causing the irregularities is still to be studied.

ACKNOWLEDGMENT

The author would like to thank colleagues from Nokia, Nokia Siemens Networks, Magister Solutions Ltd, Jyväskylä University and the Radio Network Group at Tampere University of Technology for their constructive criticism, comments and support with the work.

REFERENCES

[1] *3GPP TR 36.805, "*Study on minimization of drive-tests in Next Generation Networks*", ver. 9.0.0, December 2009.*

[2] *3GPP TS 37.320, "*Radio measurement collection for Minimization of Drive Tests*", ver. 0.7.0, June 2010.*

[3] *3GPP TR 36.902, "*Self-configuring and Self-optimization (SON) network use cases and solutions*", ver. 9.2.0, June 2010.*

[4] *R. Kreher, "*UMTS performance measurements. A Practical Guide to KPIs for the UTRAN Environment*", Wiley, 2006*

[5] *J. Laiho, A. Wacker and S Müller, "*Measurement Based Methods for WCDMA Radio Network Design Verification*", 10th Communications and Networking Simulation Symposium, March 2007.*

[6] *3GPP TS 32.421 "*Subscriber and equipment trace; Trace concepts and requirements*", ver. 9.1.0, June 2010.*

[7] *G. David and A. Averbuch, "*SpectralCAT: Categorical Spectral Clustering of Numerical and Nominal Data*", published in SampTA 2011, Singapore.*

[8] *J. Han and M. Kamber, "Data* Mining: Concepts and Techniques*", Morgan Kaufmann Publisher, 2000.*

[9] *S. Lafon, Y. Keller and R. Coifman, "*Data Fusion and Multicue Data Matching by Diffusion Maps*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, No.11, November 2006.*

[10] *G. David, A. Averbuch and R. Coifman, "*Hierarchical Clustering via Localized Diffusion Folders*", Association for the Advancement of Artificial Intelligence (AAAI), November 11-13, Virginia, 2010, USA..*

[11] *N. Rabin and A. Averbuch, "*Hierarcial Data mining approach for detection and prediction of anomalies in dynamically evolving systems*", in review process, January 2010.*

[12] *Su. Kasera and N. Narang, "*3G networks: architecture, protocols and procedures: based on 3GPP specifications for UMTS WCDMA networks*", Tata McGraw-Hill, 2004.*

[13] *A. Averbuch et al., "*Automatic discrimination between bad and good laser machines*", Applied Materials for the IMG 4 consortium, January 2010.*

A New Path Failure Detection Method for Multi-homed Transport Layer Protocol

Sinda Boussen, Nabil Tabbane and Sami Tabbane
Research Unit MEDIATRON (SUP'COM)
University of 7th November at Carthage
Tunisia
{sinda.boussen, nabil.tabbane, sami.tabbane}
@supcom.rnu.tn

Francine Krief
CNRS-LaBRI Laboratory
University of Bordeaux, IPB
France
krief@labri.fr

***Abstract*—Through its support for multi-homing, the Stream Control Transmission Protocol (SCTP) is a suitable solution to implement and manage user's mobility by abstracting multiple physical paths into a single end-to-end association. In order to detect the primary path failure, SCTP uses a strategy defined in the RFC2960 and mainly based on a retransmission time out (RTO). When a number of retransmission failures occur on the primary path, switchover procedure is initiated which means that a new primary path will be selected among the available secondary paths. In this paper, we investigate the current switchover mechanism implemented in SCTP and detail some of its deficiencies which affect the use of SCTP in a WLAN environment. Then we propose a new path failure detection strategy designed to perform path management more efficiently in wireless environment, by preempting path failure and avoiding service interruption. Finally, we outline the testing of this new strategy in the context of a WLAN environment and the results are compared to those obtained when using the standard SCTP strategy.**

Keywords- SCTP; multi-homing; RTO; path failure detection; WLAN.

I. INTRODUCTION

The Stream Control Transmission Protocol (SCTP) [7] was initially developed by the Internet Engineering Task Force (IETF) to transport signaling messages over IP networks. Compared to other transport protocols like TCP and UDP, SCTP provides additional features which are multi-homing and multi-streaming. These features make it suitable for the transport of many services which use the classical transport protocols. Currently, many applications are migrating to SCTP in order to take advantage of the new features offered by this protocol.

In SCTP terminology, an association is a connection between two endpoints which is identified by a source port and a destination port. An SCTP message contains the common SCTP header and various control or data chunks. By supporting multi-homing, SCTP is able to implement an end-to-end session transparently over multiple physical paths where the endpoint of each path is identified by an IP address. At the set up of an SCTP association, each endpoint provides a list of transport addresses composed of one or more IP addresses and a SCTP port. One of the IP addresses is used for the establishment of the primary path that is used for data chunks transmission. The other paths, called secondary paths, are used for data retransmission to increase reliability.

Moreover, through its support for multi-homing, SCTP represents a suitable solution to implement and manage user's mobility. Indeed, the primary path used for data transmission can be modified while maintaining the session. This property enables guaranteeing service continuity that is very important in some applications that rely on real time communications, such as Voice over IP (VoIP) and video streaming applications.

For that purpose, SCTP needs a path management mechanism to detect primary path failure and initiate the path switchover when necessary. The standard strategy to detect path failure, which is defined in the RFC2960, is based mainly on a retransmission Timeout (RTO). In fact, data transmission failure occurs when the timer RTO is expired without that the data sent are acquitted. Then, if the number of retransmission attempt reaches a predefined threshold called PMR, SCTP is going to activate the path switchover procedure which means that current path will be set to INACTIVE state and a new primary path will be selected.

The motivation behind this paper is a need to have a more accurate estimation of the failover (path failure) time in SCTP by interpreting the network quality degradation as an indicator of imminent primary path failure and implementing an immediate path switchover.

This paper is organized as follows. Section II details related work in the area. Section III describes in detail SCTP path management functionality. In Section IV, we propose an enhancement of the SCTP path failure detection strategy in order to preempt and avoid path failures in wireless environment. Then, Section V describes the simulated study undertaken and presents results. Finally, Section VI concludes the paper and points out future work.

II. RELATED WORK

In the current SCTP implementation, the path switchover strategy is reactive which means that switchover will only

 ISBN:978-1-61208-002-4

occur once the primary path has failed and the primary destination address is marked as INACTIVE. A number of studies have been undertaken, which investigate the performance of SCTP switchover in wireless networks.

In [1], authors show that the current SCTP mechanism for calculating RTO value is inappropriate in WLAN environments, by identifying significant deficiencies which affect the use of SCTP in a WLAN environment. These deficiencies result from the mechanism by which SCTP determines when a path switchover should be initiated. Experimental results indicate that SCTP allows more time to switchover as network conditions degrade.

In order to reduce the switchover performance deficiency experienced in WLAN environments, authors in [2] investigate the performance implications of changes to the SCTP RTO mechanism, particularly alterations to the parameters α, the smoothing factor, and β, the delay variance factor. Simulation results indicate a throughput improvement over the default mechanism defined in RFC2960, but it doesn't address the switchover delays caused by increasing RTT values in WLAN environment.

Other studies investigate how the SCTP based switchover strategies can be enhanced. In fact, a pre-emptive 802.21 oriented switchover strategy based on signal strength is proposed in [3]. According to experimental results, authors prove that the new strategy behaves more effectively than standard reactive SCTP switchover strategy, since the 802.21 standard has the ability to predict network state changes.

In [4], authors analyze the traditional failover time estimation formula in wireless networking scenarios and expose its drawbacks. Then, they propose some updates to the SCTP failover strategy in order to more accurately reflect the exact time at which primary path failure occurs.

In [5], authors propose a cross layer algorithm which uses 802.11 MAC retransmissions as an indicator of performance for all paths within an association. The use of 802.11 MAC retransmissions permit to accurately predict this performance transition significantly earlier than at the transport layer.

In [6], a cross layer approach is presented in order to manage mobility in wireless environment. It introduces local, wireless and Internet RTO subcomponents which are combined to calculate end to end RTO. It also implements a decision mechanism which selectively implements backoff on RTO subcomponents depending on network conditions.

III. CURRENT SCTP PATH MANAGEMENT

One of the features of SCTP that differentiates it from both TCP and UDP is its support of multi-homing which is the ability to support many IP addresses within an association. Multi-homing feature is used by SCTP to add resilience to network failures by providing a certain degree of network stability to critical transmission paths.

As a multi-homed protocol, SCTP needs a path management functionality to take switchover decisions as well as implementing the path switchover. To detect path failure, SCTP provides two kinds of probing mechanisms one for the primary path and another for the alternate paths. To monitor the primary path, SCTP keeps an error counter that counts the number of consecutive timeouts. For the alternate paths, SCTP uses a heartbeat mechanism to monitor the availability of these paths.

The SCTP path management functionality defines two states for each path. The state value can be set to ACTIVE or INACTIVE. A primary path is set to INACTIVE if transmission of packets on the path repeatedly fails. However, a secondary path fails, if a heartbeat chunk transmitted to the destination on that path was not successfully acknowledged. Both of these mechanisms are reactionary to network failure.

A. Path Monitoring

In SCTP associations, secondary paths are monitored to detect any changes in the reachable state of a destination address, and also to update the Round Trip Time (RTT) measurement for each of these secondary addresses. Path monitoring is performed using HEARTBEAT chunks which are sent periodically to know which addresses defined in the association are reachable (see Figure 1). When a heartbeat is received by an endpoint, the packet is processed and a heartbeat ACK packet is sent back. Each heartbeat packet contains a timestamp of when it was sent. When the heartbeat ACK packet is received, the time delay difference can be used to estimate the Round Trip Round (RTT) for secondary paths.

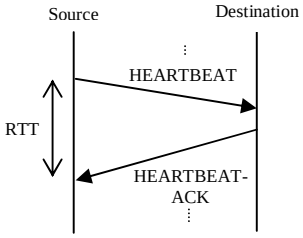


Figure 1. Secondary path monitoring

B. Retransmission Timeout Calculation

In order to detect the primary path failure, SCTP uses a reactive strategy which is mainly based on a retransmission timer. The duration of this timer is referred to as RTO (Retransmission TimeOut) [7]. The RTO duration represents the delay between each retransmission on the path. The computation and management of RTO in SCTP is similar to how TCP manages its retransmission timer. However, SCTP differs from TCP by supporting multi-homing feature. In fact, when the destination is multi-homed, the endpoint will calculate a separate RTO for each different destination's transport address.

The RTO value of the primary path is important for path switchover decision. If an SCTP sender doesn't receive a response for an SCTP data chunk from its receiver within the time of Retransmission Timeout (RTO), the sender will consider this data chunk lost. When the number of consecutive timeouts on the primary path exceeds the SCTP threshold, the address will be marked as INACTIVE by the sender, and a new primary path will be selected among the alternate paths that are currently available.

 ISBN:978-1-61208-002-4

The SCTP parameters which are used to implement the switchover management strategy are:

- RTO.Initial: the initial value for RTO.
- RTO.Min: the minimum time for RTO.
- RTO.Max: the maximum time for RTO.
- Path.Max.Retrans: the path retransmission threshold (PMR).
- HB.interval: the interval at which heartbeats are sent to monitor an SCTP endpoint.

According to [7], the following protocol parameters are recommended:

TABLE I. SCTP PARAMETERS FOR RTO CALCULATION

Parameter	Recommended Value
RTO.Initial	3 seconds
RTO.Min	1 second
RTO.Max	60 seconds
Path.Max.Retrans	5 attempts
HB.interval	30 seconds

The retransmission Timeout (RTO) is calculated for each destination address separately based on the Smoothed Round Trip Time (SRTT) and Round Trip Time Variation (RTTVAR) of the path. SRTT and RTTVAR are calculated by the measurement of Round Trip Time (RTT) of the path. Initially RTO gets RTO.initial. Then, when SCTP gets the first measurement of RTT (RTT.1st), SRTT and RTTVAR are initialized as follow:

$$SRTT = RTT.1st \quad (1)$$

$$RTTVAR = RTT.1st / 2 \quad (2)$$

And RTO is updated to:

$$RTO = SRTT + 4*RTTVAR \quad (3)$$

For each time SCTP gets a new measurement of RTT (RTT.new), SRTT and RTTVAR will be updated as follow:

$$RTTVAR.new = (1-\beta)*RTTVAR.old + \beta*(SRTT.old - RTT.new) \quad (4)$$

$$SRTT.new = (1-\alpha)*SRTT.old + \alpha*RTT.new \quad (5)$$

Where α, the smoothing factor, and β, the delay variance factor, are constants and their recommended values are 1/4 and 1/8 respectively.

Then, the new RTO is:

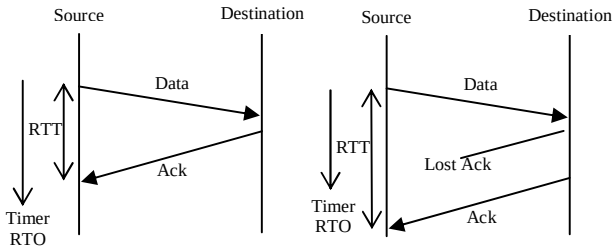
$$RTO = SRTT.new + 4*RTTVAR.new \quad (6)$$

If the new RTO is less than RTO.Min, it will be set to RTO.Min. If the new RTO is greater than RTO.Max, it will be set to RTO.Max.

Every time a transmission timeout occurs for an address (Figure 2(b)), the RTO for this address will be doubled (Backoff the time):

$$RTO = RTO \times 2 \quad (7)$$

As illustrated in Figure 2(a), if the sender gets a response from the receiver, a new RTT is measured. SCTP will use this new RTT to calculate RTTVAR, SRTT and finally RTO by the equations (4) (5) and (6).



(a) Transmission Success (b) Transmission Failure

Figure 2. Standard Path Failure Detection Strategy

C. The Standard Path Failure Detection

The standard SCTP path failure detection strategy, as illustrated in Figure 3, is based on the retransmission timer with its managing rules as defined in RFC 2960 [7].

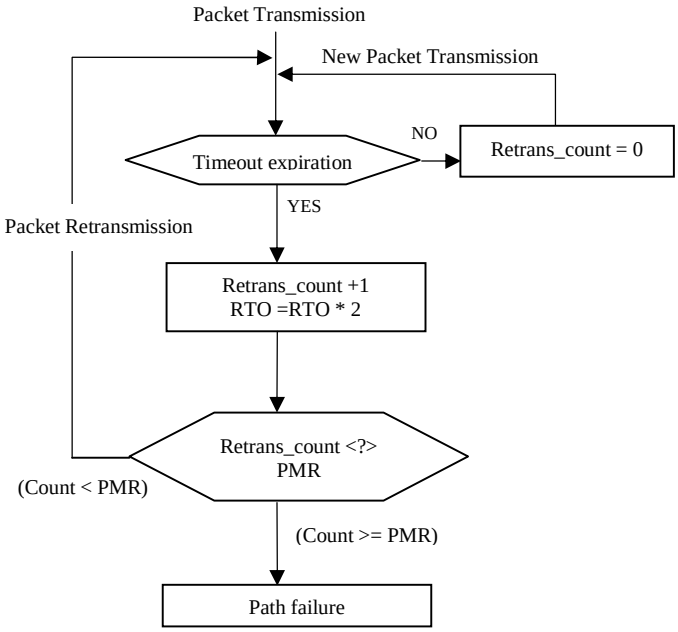


Figure 3. Standard Path Failure Detection Strategy

In fact, in SCTP association, packet transmission is through primary path only; other paths are back up in association. When a primary path is selected, the SCTP mechanism will mark the path to ACTIVE and use a retransmission count parameter to monitor path condition. If the timer expires, and the data chunk has not been acknowledged yet, it is assumed that the chunk is lost. Consequently, the actual RTO value for the affected path is doubled (exponential back-off mechanism), the error count is incremented by one and the lost chunk is marked for retransmission. When the retransmission count parameter reaches the threshold PMR (Path.Max.Retrans), the primary path takes failure. Then, the SCTP mechanism will change primary path state to INACTIVE and switch to a secondary path to continue transmission.

 ISBN:978-1-61208-002-4

IV. PROPOSED ALTERATION OF SCTP SWITCHOVER MECHANISM

Based on the sum of the consecutive retransmission timeouts, the standard strategy used by SCTP to detect the path failure is very simple and can't effectively distinguish path condition in wireless network. Consequently, this strategy is not always appropriate, especially when considering the SCTP multi-homing feature as a basis for achieving transport layer mobility in wireless network, where the transition time between available paths becomes a key aspect for the optimization.

Therefore, the most crucial challenge for SCTP is to provide optimal path management, aiming at improving the performance of the original switchover mechanism presented in Section III.

In this paper, we propose an improvement of the standard path failure detection strategy used by SCTP by changing the criteria of switchover initiation in order to obtain a more accurate estimation of the exact time at which primary path failure occurs (the Failover time). The alteration that we propose does not concern the formula of calculation of the parameter RTO. But it consist in defining new QoS parameters to preempt the path failure, and fixing thresholds to these parameters according to the type of traffic emitted and its requirements in terms of quality of service.

In fact, we propose to evaluate the Total Time spent expecting an acknowledgment (T_{ack}) in any case (packet transmission success or failure), which is an excellent indicator of path performance. T_{ack} is computed by equation (8), by representing the sum of the (k-1) consecutive timeouts according to the RTO value at the transmission failure instant. However, if the packet sent is acquitted after (k-1) retransmission attempts, T_{ack} is calculated by applying equation (9). The value (k-1) represents the number of retransmission attempts which is necessarily less than PMR ($0 < k \leq PMR$). The index j refers to the traffic type.

The time T_{ack} will be the most important parameter to consider in the SCTP switchover decision. SCTP will use it as a path performance indicator to preempt degradation in path status and avoid service interruption.

In case of (k) failed attempts

$$T_{Ack,\,j} = \sum_{i=0}^{k-1} N_j^i * RTO$$

After simplification:

$$T_{Ack,\,j} = RTO * \frac{(1 - N_j^k)}{(1 - N_j)} \qquad (8)$$

In case of success after (k-1) failed attempts:

$$T_{Ack,\,j} = RTT_{success} + \sum_{i=0}^{k-2} N_j^i * RTO$$

After simplification:

$$T_{Ack,\,j} = RTT_{success} + RTO * \frac{(1 - N_j^{k-1})}{(1 - N_j)} \qquad (9)$$

We have also introduced a new condition to evaluate the path state depending on the time T_{Ack} and the configured threshold for each type of traffic:

If $(T_{Ack,\,j} \geq T_{Threshold,\,j})$ Then (Primary path is INACTIVE)

Thus, the path is marked "INACTIVE" if one of the following conditions is satisfied:

- The number of retransmission timeouts reaches the maximum number of retransmission (PMR) authorized by SCTP.
- The waiting time T_{ack} exceeds the threshold fixed for each type of traffic (VoIP, streaming video, data)

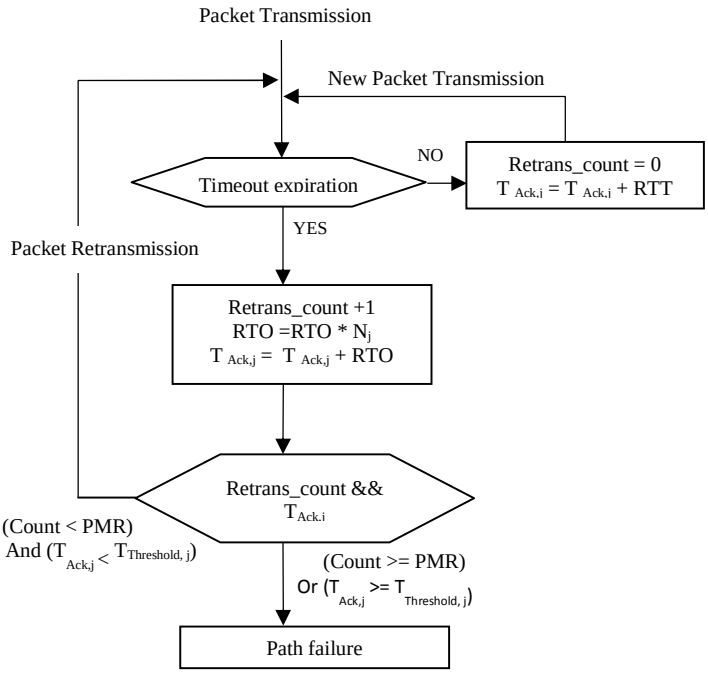


Figure 4. New Path Failure Detection Strategy

The modifications proposed with regard to the standard mechanism will be represented on the organization chart above (see Figure 4).

In RFC2960, the failure of packet retransmission induces the multiplication of the value of RTO by 2. As applications have different needs in terms of quality of service, we suggest penalizing the transmission failures in a different way for every type of traffic.

For real time streaming multimedia applications, such as voice over IP, which are delay sensitive, high latency can cause service quality degradation. However, Best effort traffic is more tolerant to delay. For these reasons, we propose to treat traffic flow differently by providing priority to certain flow, depending on their QoS requirements. We call N_j the parameter used to penalize retransmission failure, where the index j indicates the traffic type. In RFC2960, this parameter is invariable and equal to 2.

In this work, we consider the followings values:

 ISBN:978-1-61208-002-4

TABLE II. SCTP PENALIZING PARAMETERS

Traffic Type	N_i
Best Effort	2
VoIP	1
Video streaming	1.25

V. SIMULATION RESULTS

In order to illustrate the deficiencies of the strategy used by SCTP to detect the path failure and implement our proposed approach, we consider a network topology consisting of two base stations 802.11b and two nodes. The nodes are communicating and each one belongs to a base station. The network topology is shown in Figure 5.

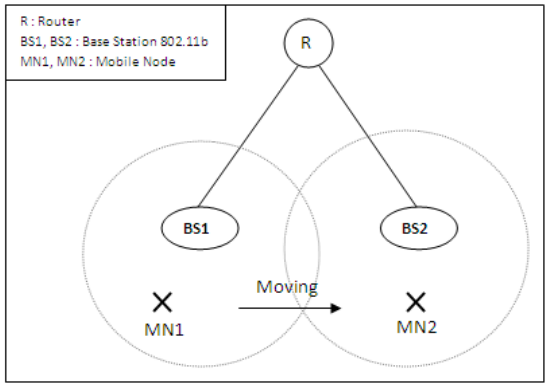


Figure 5. Simulated Network Topology

During simulation, we are interested to real-time traffics which are delay sensitive such as VoIP and video streaming. The video streaming traffic is simulated by an application that generates one packet every 26 milliseconds. Each packet has a size of 660 bytes. While the VoIP traffic is simulated by an application that generates packets of 160 bytes every 20 milliseconds. The traffic flow parameters are shown in Table 3.

TABLE III. SIMULATION TRAFFIC PARAMETERS

Traffic	Delay Interval	Packet Size	Data Rate
VoIP	20 ms	160 bytes	64 kb/s
Video	26 ms	660 bytes	200 kb/s

The simulation process time is 50 seconds, and all nodes start their transmission at 2s after the beginning of simulation time. Mobile node starts moving at 10s with a speed of 1m/s.

The simulation results presented in this paper were obtained using the network simulator NS2 [8] and the SCTP patch [9].

TABLE IV. SIMULATION PARAMETERS

Simulation Time	50 s
Trafic Start Time	2 s
Trafic Stop Time	50 s
Move Start Time	10 s
Move speed	1m/s

We will first simulate the service differentiation module of our approach which consists in penalizing transmission's failures in a different way according to traffic type.

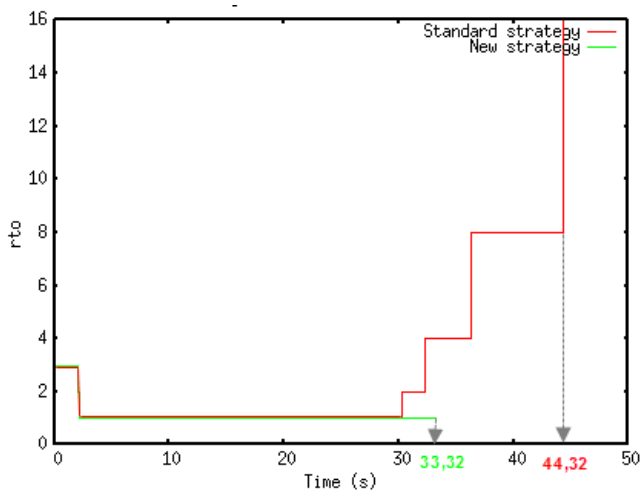


Figure 6. RTO values and Failover time detection for VoIP Traffic

Figure 6 and 7 illustrate respectively the RTO values for the simulated scenario for VoIP and video streaming traffic. When the mobile node moves away from the coverage area of the access point, signal strength degrades and the RTT and RTO increase. From simulation's result, we notice that SCTP take 15s to mark the destination address INACTIVE (T=1+2+4+8=15s). Which means that it would take 15s seconds for switchover to occur.

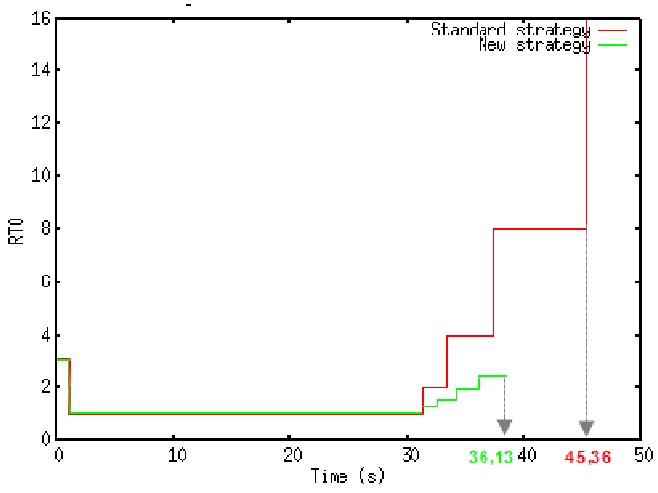


Figure 7. RTO values and Failover time detection for Video Streaming Traffic

 ISBN:978-1-61208-002-4

The failover time is defined as the instant at which the primary path failure is detected. It is computed according to RTO values and corresponds to the (PMR-1) failed attempts to retransmit a lost chunk. Table 5 represents the failover instants for VoIP and Video streaming traffic when using the standard and the new SCTP path failure detection strategy. In fact, using the standard strategy, path failover is detected at 44,32s for VoIP and 45,36s for video streaming traffic. However, when using the new strategy based on service differentiation, the path failure is detected earlier at 33,32s for VoIP and 36,13s for video streaming traffic.

TABLE V. FAILOVER TIME FOR REAL TIME TRAFFIC

Path Failure Detection Strategy	VoIP	Video Streaming
Standard strategy	44,32s	45,36s
Proposed strategy	33,32s	36,13s

In our proposed approach, we defined a second condition to detect primary path failure which is based on delay T_{ack} (Time spent expecting an acknowledgment). Figures 8(a) and 8(b) represents T_{ack} values for respectively VoIP and Video Streaming traffic. This parameter reflects the link state and therefore it can be considered to predict network performance degradation. Thus, T_{ack} will be a decisive parameter to initiate switchover process.

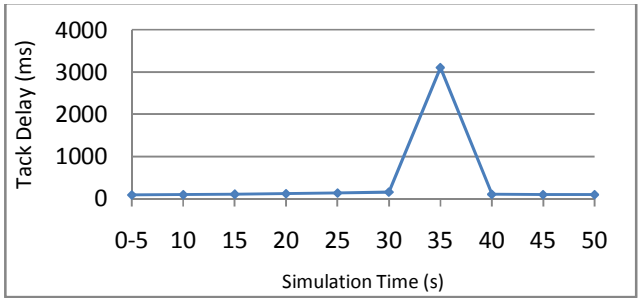


Figure 8-(a) T_{ack} Delay for VoIP Traffic

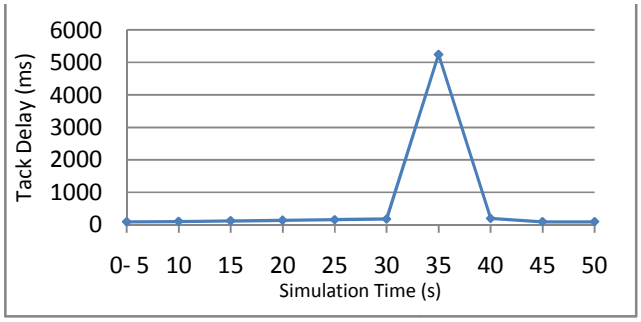


Figure 8-(b) T_{ack} Delay for Video Streaming Traffic

Figure 8. T_{ack} Delay for for Real Time Traffic

To further illustrate the shortcomings of the standard method and highlight the contribution of our new approach to detect primary path failure, we will represent network's performance metrics such as throughput, delay and loss.

1) Throughput

The throughput, measured in kbps, corresponds to the amount of data in bits that is transmitted over the channel per unit time.

$$Throughput = \frac{Total\ number\ of\ bits\ successfully\ transmitted\ during\ T}{T}$$

2) End-to-End Delay

The end to end delay, measured in second, is the time taken for a packet to be transmitted across a network from source to destination. It is an important parameter to evaluate the QoS for the real-time traffic.

$$Delay = \frac{\sum_{i=0}^{N} (Time\ of\ packet_i\ received\ -\ Time\ of\ packet_i\ sent)}{Total\ number\ of\ packet\ received}$$

3) Packet Loss Rate

Packet loss is expressed as a percentage of the number of packets lost to the total number of packets sent.

$$Packet\ Loss\ Rate = \frac{Number\ of\ dropped\ data\ packet}{Total\ number\ of\ packet\ data\ sent} \times 100$$

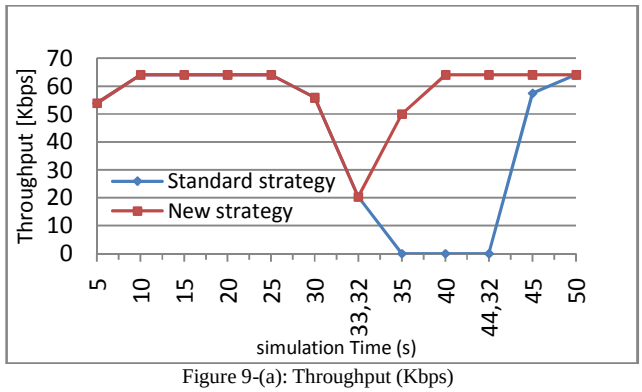


Figure 9-(a): Throughput (Kbps)

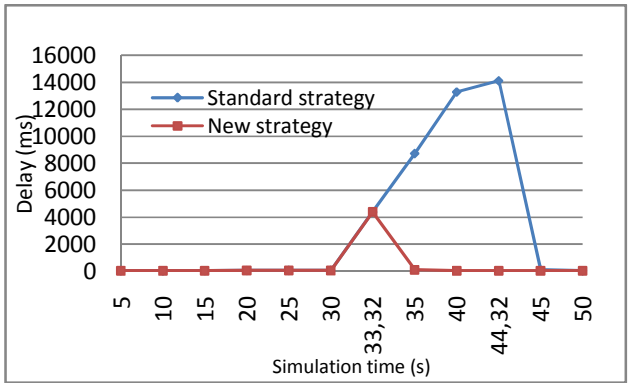


Figure 9-(b): End To End Delay (ms)

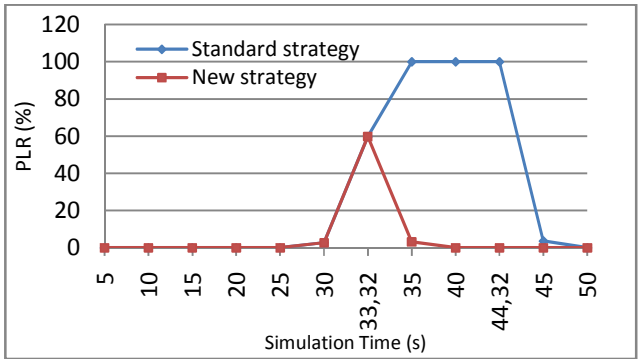


Figure 9-(c): Packet Loss Rate

Figure 9. Performance metrics for VoIP traffic

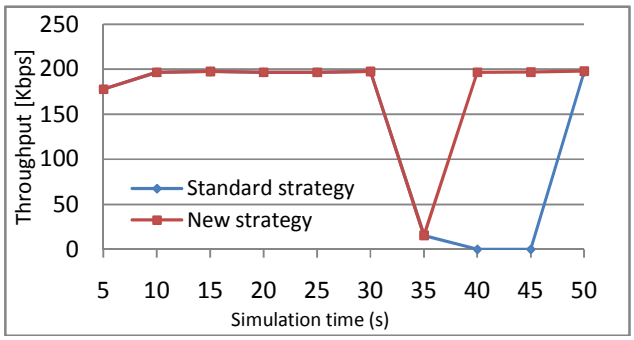


Figure 10-(a): Throughput (Kbps)

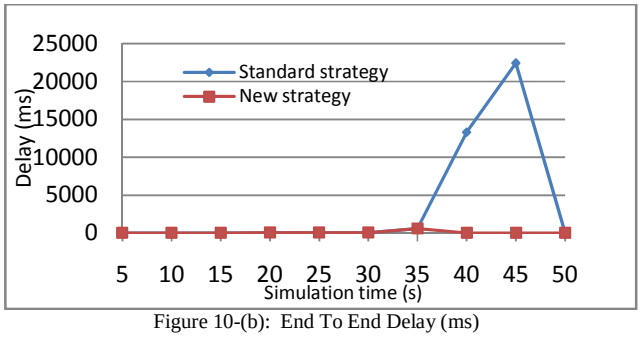


Figure 10-(b): End To End Delay (ms)

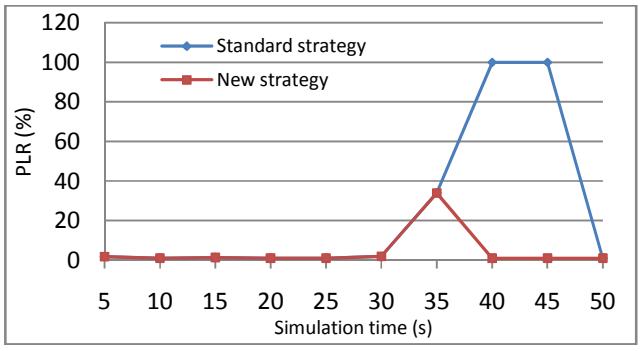


Figure 10-(c): Packet Loss Rate

Figure 10. Performance metrics for Video Streaming traffic

Figures 9 and 10 represent the metrics of network performance (Throughput, Delay and Packet loss) for respectively VoIP and Video streaming traffic. According to simulation's result, we notice that when using the standard strategy, although there was a degradation of network performance in terms of throughput, delay and loss, SCTP delays switchover, i.e., SCTP allows more time to initiate switchover.

Through simulation results, we deduce that our approach could be an alternative to the current SCTP path failure detection strategy used by SCTP; by increasing network performance and providing a seamless switchover to real-time applications such as VoIP or video streaming

VI. CONCLUSION

In this paper, we are interested to the mechanism used by SCTP to take the decision of changing the primary path which relies mainly on the failover mechanism. We have detailed the current mechanism implemented in SCTP and described some of its failings. Then, we have proposed a proactive approach to detect the path failure. Our approach would be more suitable to a mobile environment such as WLAN. In fact, according to experiment results, the proposed approach allows SCTP to detect the path failover earlier than the standard mechanism. Moreover, it provides a seamless switchover to real-time applications by increasing network performance and avoiding service interruption. In future work, we will investigate the algorithm used by SCTP to estimate the RTO timer, in order to enhance switchover performance in WLAN environment.

REFERENCES

[1] S. Fallon, P. Jacob, Y. Qiao, and L. Murphy, "SCTP Switchover Performance Issues in WLAN Environments", 5th IEEE Consumer Communications and Networking Conference (CCNC) 2008, Issue 10- 12 Jan. 2008, pp. 564 – 568

[2] S. Fallon, P. Jacob, Y. Qiao, and L. Murphy, "An Analysis of Alterations to the SCTP RTO Calculation Mechanism for WLAN Environments". In MWCN 2008 Wireless and Mobile Networking. IFIP, vol. 284, pp. 95–108 (2008)

[3] E. Fallon, J. Murphy, L. Murphy, Y. Qiao, X. Xie, and A. Hanley, " Towards a Media Independent Handover Based Approch to Heterogenous Network Mobility", In Proceedings of The IET Irish Signals and Systems Conference 2007 (ISSC07).

[4] L. Budzisz, R. Ferrus, K. Grinnemo, A. Brunstrom, and C. Ferran, "An Analytical Estimation of the Failover Time in SCTP Multihoming Scenarios", Wireless Communications and Networking Conference (WCNC) 2007.

[5] S. Fallon, P. Jacob, Y. Qiao, A. Hanley, and L. Murphy, "Using 802.11 MAC Retransmissions for Path Selection in Multi-homed Transport Layer Protocols". In Proceedings of the IEEE Global Communications Conference (GLOBECOM 09).

[6] S. Fallon, P. Jacob, Y. Qiao, and L. Murphy, "An Adaptive Optimized RTO Algorithm for Multi-homed Wireless Environments", 7th International Conference on Wired and Wireless Communications (WWIC) LNCS 5546, pp. 133-145 (2009)

[7] R. Stewart, Q. Xie, K. Morneault, C. Sharp, and H. Schwarzbauer, T.Taylor, I.Rytina, M.Kalla, L.Zhang, V. Paxson. "RFC2960 - Stream Control Transmission Protocol ".October 2000.

[8] Network Simulator- NS2- http://www.isi.edu/nsnam/ns.

[9] SCTP Patch for NS. http://pel.cis.udel.edu.

 ISBN:978-1-61208-002-4

A Mechanism for Semantic Web Services Discovery in Mobile Environments

Rafael Besen, Frank Siqueira
Department of Informatics and Statistics (INE)
Federal University of Santa Catarina (UFSC)
Florianópolis, Brazil
{rbesen, frank}@inf.ufsc.br

***Abstract*—This paper describes SeMoSD (Semantic Mobile Service Discovery), a mechanism for discovery of services in mobile environments based on web services technology. The service discovery mechanism was designed taking into account the resource constraints of mobile devices. Through SeMoSD, mobile devices are able to profit from the use of semantic technologies for service discovery. Services are described using ontologies, and may be executed on either mobile or fixed nodes. Queries for services are interpreted by a semantic reasoner, resulting in more accurate search results.**

Keywords- Semantic Web Services, Ontology, Mobile Devices, DPWS, WSMO, Ubiquitous Computing.

I. INTRODUCTION

Computing resources provided by mobile devices may be shared with other nodes and employed to perform computing tasks cooperatively. In order to allow interoperation in a distributed environment, mobile devices must be able to locate each other on the network. Therefore, mechanisms for discovery of networked resources play an important role in the construction of distributed application in mobile environments.

The first step in the discovery process is the description of the shared resource. One widely adopted approach for sharing networked resources is the Service-Oriented Architecture (SOA) [1], which models services provided by computing devices that can have their execution requested remotely.

Services are usually described syntactically, through the specification of service names (i.e. a method or procedure) and messages (i.e. parameters). A purely syntactic description is a very simple and limited way to do this, because the same identifiers employed to describe the service upon registration must be specified when the service is queried.

On the other hand, a semantic description of services allows a richer amount of information regarding the provided services to be specified. Such technology is starting to be employed for the description and location of networked services, but is still novelty in the context of mobile computing devices.

This work describes SeMoSD (Semantic Mobile Service Discovery), a service discovery mechanism that aims to bring the benefits of semantic technologies to mobile computing environments. This is achieved through the definition of a model for interaction among devices, which is based on the DPWS (Devices Profile for Web Services) standard [2]. Semantic description and query processing rely on WSMO (Web Services Modeling Ontology) [3] and its execution engine WSMX (Web Services Modeling eXecution environment) [4].

The remainder of this paper is organized as follows. Section II analyzes the use of web services technology in the context of an environment composed by mobile devices. Semantic technologies adopted in this work are described in Section III. Section IV presents the architectural characteristics and execution dynamics of the proposed semantic discovery mechanism. An application scenario which employs the proposed discovery mechanism is described in Section V. Section VI presents some related research projects and compares them with the solution proposed in this paper. Finally, Section VII sums up the contributions given by this work and suggests further developments in this field.

II. MOBILE DEVICES AND WEB SERVICES

Mobile computing devices, such as smartphones, media players, tablets, wireless sensors and RFID tags, become more and more common every day. However, allowing these devices to interoperate seamlessly and to execute management tasks autonomously is still a distant reality. Devices from different vendors often employ different communication technologies, making them unlikely to be able to identify each other and exchange data in order to work together. In addition, in the context of mobile devices, care must be taken with some issues, such as frequent network disconnection, reduced processing power and storage capacity, and limited battery life.

Since these devices may join and leave the network at any time, requiring system reconfiguration, devices depend even more on the discovery mechanisms for locating new partners for the execution of cooperative tasks and for restoring dependencies on remote services.

The currently adopted mechanisms for device discovery in mobile environments are mostly based on syntactic information. Bluetooth SDP [5], Jini [6], UPnP [7] and Salutation [8] have discovery mechanisms with this limitation.

Attempts have been made to employ the Web Services technology [9] in the context of mobile devices. This technology has been widely adopted for building distributed

 ISBN:978-1-61208-113-7

applications and for integrating legacy software. One of the main reasons for the success of this technology is the adoption of a set of broadly available standards – such as the eXtensible Markup Language (XML) [10], and the SOAP [11] and HTTP [12] communication protocols. This fact allows developers to create web services using several different programming languages and operating systems, in a truly heterogeneous environment.

However, the lack of appropriate mechanisms for description and discovery of services is still a limiting factor for the use of this technology. The Web Services Description Language (WSDL) [13] and the Universal Description, Discovery and Integration (UDDI) [14] provide means only for syntactic description and discovery of services, while semantic information on services is required for locating services without human intervention. Furthermore, these standards do not provide the required flexibility for describing and discovering services in a constantly changing environment, such as an ad-hoc network composed by mobile devices.

A. DPWS

The limitations for hosting web services in mobile devices are partially solved with the use of a standard called DPWS (The Devices Profile for Web Services) [2]. DPWS is a standard published by OASIS, which defines a communication model based on the Web Services technology, tailored for use in mobile computing devices. The adopted strategy allows devices connected to a network to syntactically describe the services provided by them and advertise these services to other devices. Devices can then have their services discovered dynamically and invoked by other devices which are able to understand the syntactic description of the service. Any device with connectivity and processing resources – such as mobile computers, smartphones, tablets, sensors and many others – can implement this standard and access or provide services.

DPWS identifies two kinds of services: hosting services, which represent devices, provide data about themselves to allow device discovery; and hosted services, which are computational services hosted by hosting services. They have their own network addresses in order to be reachable by other devices.

The flow of DPWS messages between a client device and a hosting service (device) with one hosted service is shown by Fig. 1. A client device broadcasts a *Probe* request, expressing its will to find other devices with characteristics specified in the message body. A device that matches the request replies with a *Probe Match* message. After getting in touch with one or more devices, the client may send the *Get Metadata* message to receive more information related to the device, including the list of available services. With this knowledge in hand, the client can send *Get Metadata* messages directly to the hosted service to get the required information for invoking the service. After receiving this information, the device can finally invoke the service.

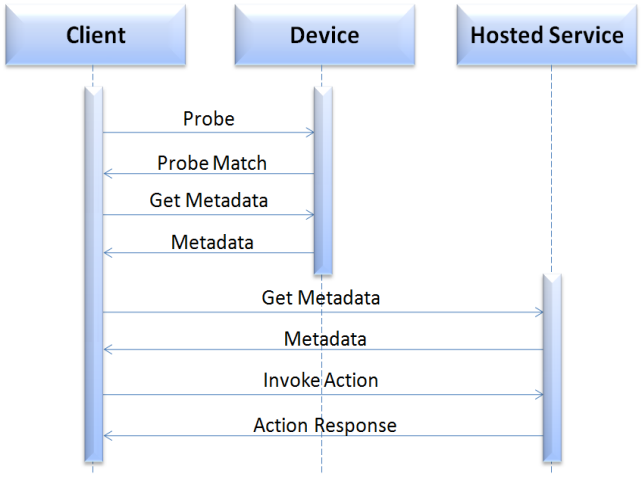


Figure 1. DPWS Messages

III. SEMANTIC TECHNOLOGIES

Better accuracy in service location may be obtained through the use of semantic technologies, which are based on the use of ontologies. Services described using ontologies are the base of a technology called Semantic Web Services [15].

An ontology can describe a set of interrelated concepts which belong to a certain domain, allowing these concepts to be associated to service operations and data exchanged with services. The use of ontologies allows searches to produce more accurate results, since it allows the exact meaning of data and the search context to be precisely specified. Furthermore, the described concepts are machine-readable, reducing the necessity of human intervention in the service discovery process.

Due to the resource constraints presented by mobile devices, the execution of semantic searches and the storage of semantic data may need resources that most devices are unable to provide. On the other hand, the provision of semantic web services can make feasible the creation of a ubiquitous computing environment [16], using the description of services to integrate them, and making a better use of resources provided by mobile devices.

The following sections present, respectively, the semantic language and the execution engine adopted in this work for semantic description of services hosted by mobile devices. These were chosen due to their superior characteristics and expressiveness over other semantic technologies, such as OWL-S [17].

A. WSMO

The Web Service Modeling Ontology [3] is a promising technology for developing Semantic Web Services. The interactions and data exchanged with a service are described using ontologies written in WSML (Web Services Modeling Language) [18]. The low-level representation of this language is based on XML, and shared resources are identified by URIs (Universal Resource Identifiers).

 ISBN:978-1-61208-113-7

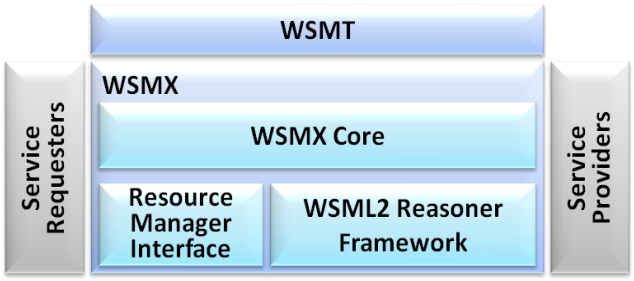


Figure 2. WSMX Architecture

WSMO allows roles and concepts to be expressed separately. Client requirements and available services are specified in different documents. Service description and implementation are also defined separately. This strategy improves reuse and maintains backward compatibility with systems that lack semantic capabilities.

The description and discovery of services is based on ontologies, which must define the meaning of data exchanged through messages. Services are described using concepts, relations and instances of this ontology. To identify a desired service, a goal is specified, containing all features needed in a service. Semantic reasoners compare goals with the description of the available services in order to find one or more matching services.

B. WSMX

The Web Services Modeling eXecution environment [4] is an execution engine for WSMO. WSMX allows services to be integrated in an automatic, flexible and easy way. The WSMX engine allows developers to integrate legacy systems without requiring internal changes in their source code.

The WSMX is composed by a set of components, each one with a service interface and a specific role. These components, which are illustrated by Fig. 2, include a reasoner framework [19][17], which allows the user to choose or add semantic matching algorithms; and WSMT [20], which is a tool that helps users to manage semantic descriptions and to interact with the execution engine.

IV. SEMOSD – SEMANTIC MOBILE SERVICE DISCOVERY

The Semantic Mobile Service Discovery infrastructure is aimed at providing a repository for storing and locating semantically-rich service descriptions in an environment composed by mobile devices. Such a repository is necessary because a typical mobile device does not have enough computing resources for storing semantic service descriptions in a local cache and also for executing the semantic reasoning algorithms.

The repository implementation and all the interaction with it are based on widely available standards and technologies, which had to be adapted and integrated to allow the execution of semantic services on mobile devices. In the center of this infrastructure is an implementation of the DPWS standard, which provides the communication mechanisms for interaction among mobile devices.

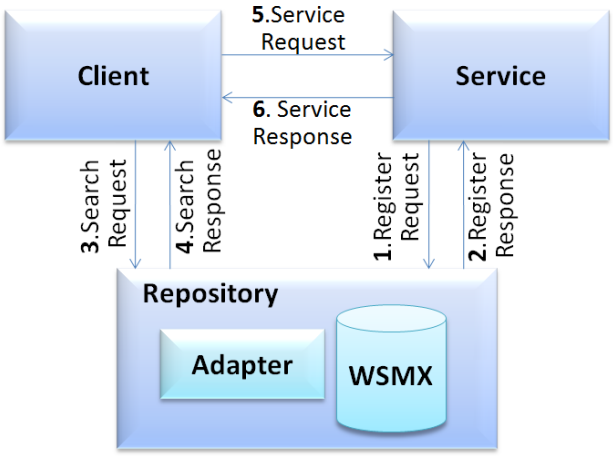


Figure 3. Execution Model

Metadata obtained through DPWS holds service descriptions. Since the standard does not define how services are described, it is possible provide a semantic description for a service as metadata. To maintain backward compatibility with already existing DPWS applications, which are unable to understand a semantic description, a syntactic description is maintained in this attribute. This is done by extending the metadata class to carry also semantic data, keeping the syntactic data untouched. Therefore, clients that are unable to interpret the semantic description may participate in a regular DPWS client-server interaction.

Ontologies must be employed to semantically describe services, defining device concepts, services, relationships and instances. In this work we have adopted WSMO, which was described in Section **Erro! Fonte de referência não encontrada.**. Services described semantically (i.e., using concepts and relationships defined in an ontology) may be more easily and precisely located by clients. The characteristics of the desired service may be described by the client building a goal using WSML syntax. Alternatively, a client may just specify parameter values to specialize a goal stored in the semantic repository. After locating one or more service descriptions in the repository, the client can choose one or more services to interact with, issuing regular service invocations.

The semantic repository employs an adapter, which receives DPWS requests from client devices, parses these requests and sends them to WSMX. The adapter must implement the DPWS protocol stack and run on the same platform as the WSMX engine. Reasoners provided by WSMX are responsible for locating services that match the characteristics described in the goal executed by the client device. The adapter waits for a response from the reasoner and, as soon a list of matching services becomes available, a response is given to the client device.

Since the reasoning algorithms require more processing power than is currently available in most mobile devices, and taking into account the importance of this service in the environment, the semantic repository might have to be hosted by a device with more processing power and storage capacity and with non-intermittent network connectivity.

 ISBN:978-1-61208-113-7

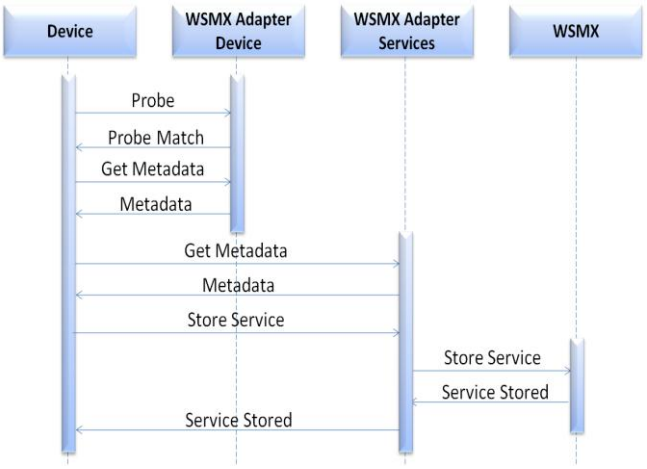


Figure 4. Service Behaviour

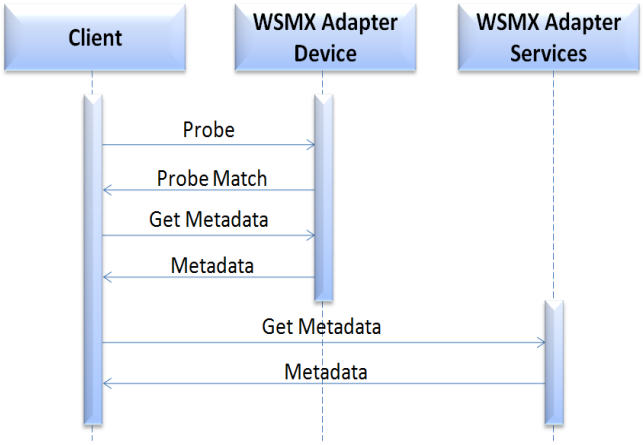


Figure 5. Client Behaviour

A. *Execution Model*

Fig. 3 explains the interaction among devices during the publication of a service description and the execution of a search for the described service. Upon startup, a device – either service or client – must broadcast a DPWS probe in an attempt to locate a semantic discovery mechanism within its reach. Thus, the first request is a regular DPWS request, which does not require any semantic knowledge. To allow the semantic discovery mechanism to be located through a regular DPWS request, it must always be described in DPWS Metadata as "WSMX Service".

If there is no semantic discovery mechanism available, the device tries to locate services using the discovery mechanism provided by the DPWS protocol. When a semantic discovery mechanism connects to the network, all services are notified through an announcement message and, from this moment on, they may start using the mechanism – i.e. they may register their services and ask for the location of services described semantically.

Fig. 4 explains in more detail the request flow between a service provider and the SeMoSD infrastructure. When a provider connects to the network, it broadcasts a DPWS probe message, willing to find a semantic discovery mechanism. After identifying a repository (*Probe Match* message), the provider can obtain metadata from the device and from the WSMX service (*Get Metadata* and *Metadata* message), and store the description of its services (*Store Service* message, confirmed with a *Service Stored* message).

The same startup procedure is executed by clients to locate the WSMX Service. The exchanged messages are illustrated by Fig. 5. After this initial step, clients may specify goals that will be executed by the WSMX Service, in an attempt to locate a required service. New goals can be executed, and preexisting goals can be parameterized. When a request is received, the WSMX Adapter parses its contents and sends the request to WSMX. Then, the semantic reasoning is executed and the response is sent to the adapter, which forwards the response to the client. With the search results in hand, the client can invoke these services directly using the DPWS protocol.

B. *Prototype Implementation*

A prototype of the SeMoSD infrastructure was developed in Java, using JDK version 6. The Java Platform was chosen due to the extensive support for mobile devices and the easy integration with WSMX. Version 0.5 of WSMX was adopted in this prototype, due to the fact that the 1.0 version is still in beta release and is not stable. The adapter, which allows the integration between DPWS and WSMX, translates requests into WSML using the WSMX Integration API.

Clients and services for testing purposes were developed using the JMEDS implementation of DPWS, but any other DPWS implementation can be adopted for this purpose. WSMO4J was employed for describing services and goals in WSML.

Performance measurements obtained during tests with this prototype have shown that searches executed through the adapter, which was co-located with the WSMX service, took approximately 139 ms longer than a direct request to the WSMX service. The total search time in a WSMX service with a minimal set of registered services takes 3.59 seconds – i.e., the overhead imposed by the proposed mechanisms is lower than 4%.

V. CASE STUDY

Fig. 6 shows a usage scenario chosen to illustrate the use of the SeMoSD infrastructure. This scenario consists in the computing support for a disaster relief operation in a site where, due to natural phenomena such as an earthquake, a mudslide, intense flooding or a hurricane, several people, either injured or with their lives at risk, must be rescued as soon as possible.

When multiple rescue teams arrive at the location, they must be aware of the resources available at the site and in the neighborhood. Besides, it is necessary to provide information about the rescue to families and to the press.

The semantic infrastructure can be useful to allow the cooperation among rescue teams and to coordinate the use of human and physical resources, such as medical staff, firemen, truck drivers, and also ambulances, bulldozers, fire trucks, helicopters, and so on.

 ISBN:978-1-61208-113-7

Figure 6. Rescue Operation Scenario

Suppose that each city or state has a repository that gathers information about all the local resources available for a disaster relief operation, which are described using the ontology depicted in Fig. 7. Information systems and mapping services are also necessary for executing the disaster relief operation. With such information available, teams can work cooperatively, coordinating the use of local resources and of services provided by all teams that take part in the operation.

Upon arrival at the site, rescue personnel must connect their mobile devices to the local semantic repository, and identify the services available. Each team must publicize a semantic service providing information about the services it provides, such as transportation, healthcare and so forth. Available equipment with network connectivity may also provide real-time information about their location and availability.

For example, a team rescuing people from a collapsed building may request assistance from first-aid teams to provide preliminary medical care. These can locate ambulances to take the injured to the closest hospital with available beds, equipment and personnel to give the required treatment. To do so, first-aid teams and ambulances must provide their GPS coordinates and current availability. Hospitals must also provide real-time information regarding the available medical services. Based on this information, cooperation strategies may be more precisely and efficiently architected and executed.

VI. RELATED WORK

Most of the existing discovery mechanisms for mobile environments do not provide support for semantic discovery of devices and services. Some technologies that fall in this category are already in use, such as Bluetooth SDP, Jini and UPnP. Others, such as TOTA Approach [21] and the Device Service Bus [22], have been proposed recently and are still experimental work. In dynamic environments with multiple heterogeneous devices, it is very unlikely that devices will be able to interpret syntactic data without human intervention.

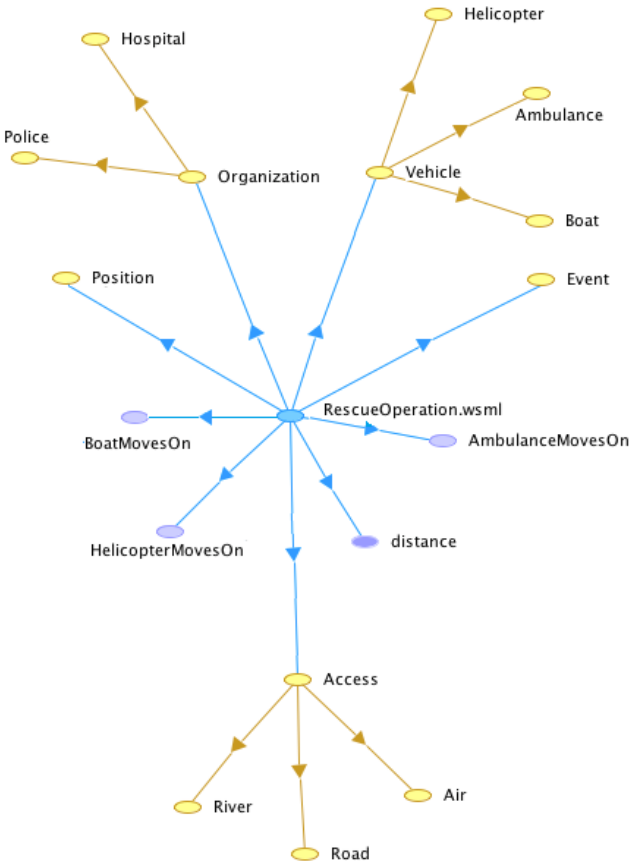


Figure 7. Ontology for Rescue Operations

Some applications employ semantic data to describe services, such as AIDAS [23], which is developed in Java and uses information from mobile devices to catch user context information. Based on this information, semantic reasoning is performed to find the available services. However, the mobile devices are employed only to retrieve information, and the proposal is limited to the Java language.

Another technology that employs semantic knowledge to locate services is Home-SOA [24]. This project uses OSGi components, because the authors believe that web services are not the best choice to do the communication, due to the high traffic created. Home-SOA defines communication interfaces that are limited to languages supported by the Java Virtual Machine, and its use is also limited to small domestic networks, composed mainly by multimedia devices.

The GEA architecture [25] is targeted at government services. To describe services, GEA uses ontologies written in OWL-S [17]. Services in this architecture are mainly to process documents and to access document models and sheets. The use of mobile devices is not addressed by this architecture.

Table 1 presents a comparison between these projects and the semantic discovery service proposed in this paper. Other projects in this area also have limitations regarding the supported programming languages, the lack of adherence to open standards, the use of restrictive environments, or the lack of support for mobile devices.

 ISBN:978-1-61208-113-7

TABLE I. CHARACTERISTICS OF SEMANTIC MODELS

Feature	Semantic Models			
	AIDAS	***Home-SOA***	***GEA***	***SeMoSD***
Communi-cation	Web Services	OSGi	Web Services	Web Services (DPWS)
Language Constraints	Java	JVM	None	None
Scope	Any	Home Media	Electronic documents	Any
Mobility Support	Limited	Yes	None	Yes

VII. FINAL REMARKS

This paper described SeMoSD, a semantic discovery infrastructure aimed at reaching a balance between providing accuracy of service location and adequacy for mobile environments. Its adoption simplifies the construction of flexible ubiquitous computing environments.

The proposed infrastructure is based on standards which, combined, allow the semantic location of services hosted by mobile devices. The amount of network messages necessary to locate a service is reduced comparing to other location mechanisms, avoiding the excessive use of battery, which limits the lifetime of devices.

A prototype of the proposed mechanism was implemented to verify its feasibility, and tests with this prototype have shown that it introduces a small overhead (lower than 4%) in the total search time. As future work, we intend to test the prototype in a real-life application in order to validate the proposed mechanism.

REFERENCES

[1] T. Erl, "Service-Oriented Architecture: Concepts, Technology and Design", Prentice Hall, 2005.

[2] OASIS, "OASIS Devices Profile for Web Services (DPWS) Version 1.1", July 2009, http://docs.oasis-open.org/ws-dd/ns/dpws/2009/01, Accessed: September 25, 2010.

[3] ESSI WSMO working group, "D2v1.3. Web Service Modeling Ontology (WSMO)", October 2006, http://www.wsmo.org/TR/d2/v1.3/, Accessed: September 25, 2010.

[4] T. Hasselwanter, P. Kotinurmi, M. Moran, T. Vitvar, and M. Zaremba, "WSMX: a Semantic Service Oriented Middleware for B2B Integration", Proc. 4th Intl. Conf. on Service Oriented Computing, pp. 477-483, Chicago, USA, December 2006, DOI: 10.1007/11948148_43.

[5] BlueTooth SIG, "BlueTooth Specification v. 3.0", April 2009, http://www.bluetooth.com/Bluetooth/Technology/Building/ Specifications, Accessed: September 25, 2010.

[6] Sun Microsystems, "Jini Technology Core Platform Specification v2.0", June 2003.

[7] UPnP Forum, "UPnP Device Architecture 1.0", December 2003.

[8] Salutation Consortium, "Salutation Architecture Specification", 1999.

[9] W3C, "Web Services Architecture", February 2004, http://www.w3.org/TR/ws-arch/, Accessed: September 25, 2010.

[10] W3C, "Extensible Markup Language (XML)", September 2003, http://www.w3.org/XML/, Accessed: September 25, 2010.

[11] W3C, "SOAP Version 1.2 Part 1: Messaging Framework", April 2007, http://www.w3.org/TR/soap12-part1/, Accessed: September 25, 2010.

[12] R. Fielding et al., "Hypertext Transfer Protocol – HTTP/1.1", IETF RFC 2616, June 1999, http://www.w3.org/Protocols/ rfc2616/rfc2616.html, Accessed: September 25, 2010.

[13] W3C, "Web Services Description Language (WSDL) 1.1", March 2001, http://www.w3.org/TR/wsdl, Accessed: September 25, 2010.

[14] OASIS, "UDDI Version 3.0.2", October 2004, http://www.uddi.org/pubs/uddi_v3.htm, Accessed: September 25, 2010.

[15] J. Cardoso, "Semantic Web Services: Theory, Tools and Applications", Information Science Reference, 2006.

[16] M. Weiser, "Hot Topics: Ubiquitous Computing", IEEE Computer, 26(10), pp. 71– 72, October 1993, DOI: 10.1109/2.237456.

[17] W3C, "OWL-S: Semantic Markup for Web Services", November 2004, http://www.w3.org/Submission/OWL-S/, Accessed: September 25, 2010.

[18] ESSI WSMO working group, "The Web Service Modeling Language WSML", August 2008, http://www.wsmo.org/wsml /wsml-syntax, Accessed: September 25, 2010.

[19] S. Grimm, U. Keller, H. Lausen, and G. Nagypal, "A Reasoning Framework for Rule-Based WSML", Proc. 4th European Semantic Web Conference (ESWC), Innsbruck, Austria, June 2007, pp. 114-128, DOI: 10.1007/978-3-540-72667-8_10.

[20] M. Kerrigan, "The Web Services Modelling Toolkit (WSMT)", Technical report, 2005, http://www.wsmo.org/ TR/d9/d9.1/v0.2/, Accessed: September 25, 2010.

[21] M. Mamei and F. Zambonelli, "Programming pervasive and mobile computing applications: The TOTA approach", ACM Trans. Softw. Eng. Methodol., 18(4), July 2009, DOI: 10.1145/1538942.1538945.

[22] G. M. Araújo and F. Siqueira, "The Device Service Bus: A Solution for Embedded Device Integration through Web Services", 24th ACM Symposium on Applied Computing, Honolulu, Hawaii, USA, March 2009, pp. 185-189, DOI: 10.1145/1529282.1529322.

[23] A. Toninelli, A. Corradi, and R. Montanari, "Semantic-based discovery to support mobile context-aware service access", Computer Comunication, 31: pp. 935-949, January 2008, DOI:10.1016/j.comcom.2007.12.026.

[24] A. Bottaro and A. Gérodolle, "Home SOA – Facing Protocol Heterogeneity in Pervasive Application", Proc. Intl. Conf. Pervasive Services, Sorrento, Italy, July 2008, pp. 73-80, DOI: 10.1145/1387269.1387284.

[25] V. Peristeras and K. Tarabanis, "Advancing the Government Enterprise Architecture – GEA: The Service Execution Object Model", R. Traunmüller (Ed.), Electronic Government 2004, pp. 476–482, 2004, DOI: 10.1007/978-3-540-30078-6_83.

 ISBN:978-1-61208-113-7

Towards Efficient Energy Management: Defining HEMS, AMI and Smart Grid Objectives

Ana Rosselló-Busquet, Georgios Kardaras, José Soler and Lars Dittmann, *IEEE members*
Networks Technology & Service Platforms group, Department of Photonics Engineering,
Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
{*aros, geka, joss, ladit*}*@fotonik.dtu.dk*

***Abstract*—Energy consumption has increased considerably in the recent years. The way to reduce and make energy consumption more efficient has become of great interest for researchers. One of the research areas is the reduction of energy consumption in users' residences. Efficiently managing and distributing electricity in the grid will also help to reduce the increase of energy consumption in the future. In order to reduce energy consumption in home environments, researches have been designing Home Energy Management Systems (HEMS). In addition, Advanced Metering Infrastructure (AMI) and smart grids are also being developed to distribute and produce electricity efficiently. This paper presents the high level goals and requirements of HEMS. Additionally, it gives an overview of Advanced Metering Infrastructure benefits and smart grids objectives.**

I. Introduction

Despite the fact that home appliances have become more energy efficient, electricity consumption in households has increased 30% over the last 30 years [1]. This is due to the fact that the number of appliances that can be found in households is also increasing. According to the International Energy Agency (IEA), European electricity consumption is going to increase 1.4% per year up to 2030 unless countermeasures are taken [2].

Residential buildings can reduce their energy consumption by becoming more energy efficient. This paper will try to identify the objectives that need to be fulfilled in order to deploy an energy efficient infrastructure. This infrastructure will help reduce the electricity consumption in users' residencies and make the electric grid more efficient.

The research areas of efficient energy management have divided into three more specific research areas: energy management in-home environments, consumers and utilities cooperation and energy management in the electrical grid.

In this paper we treat 'utilities' as the parties involved in the production and distribution of electricity through the electrical grid. In addition, we use the term distribution to refer to the process of electricity transport from the generation plants to the users' residences.

As shown in table I, energy consumption in home environments can be reduced by installing Home Energy Management System (HEMS) [3] in users' residences. HEMS will give the users the necessary tools to manage and reduce their consumption. Advanced Metering Information (AMI) will enable two way communication between the households and the utilities. AMI will benefit the users as it will enable the provision of real time rates and billing status through the smart meter. If users take into consideration the price of electricity while consuming and reduce their consumption when the price is high, consumption will be optimized as demand peaks will be reduced. In addition, providing this exchange of information is one of the first steps towards optimization of energy distribution and production as it will provide the utilities with statistics that will help predict energy consumption. In order to reduce losses and optimize energy distribution and production the electrical grid needs to be upgraded. Upgrading the electrical grid will lead to the so called smart grid. The smart grid will include new elements to efficiently manage the electricity distribution and production.

In this paper, the different goals that should be achieved in these areas in order to reduce energy consumption in home environment and make more efficient distribution and production of electricity are presented. When designing such systems, researchers usually focus on one of the goals. However, it is important that when designing these systems, researchers design them in the framework they are going to be deployed and keep in mind all the goals they should achieve to maximize the benefits. This paper summarizes the different objectives of these research areas which can be used as a guideline.

The remainder of this paper is organized as follows: Section II introduces the concept of Home Energy Management System (HEMS) and describes the high level goals and requirements to deploy it successfully. Section III will present the concept of smart meters and AMI. In addition, the possible information exchange between households and utilities will be explained. Finally, section IV will present the concept of smart grid and the objectives that need to be achieved to optimize energy production and distribution.

II. Energy management in home environments

Energy consumption in households should be reduced in order to decrease greenhouse gas emissions. Introducing Information and Communication Technologies (ICT) into home environments can help reduce users' energy consumption. A HEMS is a system that includes all the necessary elements to achieve reduction of electricity consumption in home environments. One of its main elements is the so

 ISBN:978-1-61208-113-7

Table I
IMPROVING ENERGY MANAGEMENT

Issues \ Research areas	Home environments	Cooperation	Electrical Grid
Energy Goals	Reduce energy consumption	Optimize energy consumption, distribution and production	Reduce losses and optimize energy distribution and production
Who benefits?	Users	Users and utilities	Utilities
How?	HEMS	AMI	Smart grid

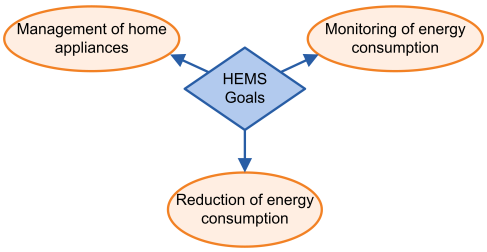


Figure 1. Home Energy Management System Goals

called home gateway or residential gateway which is able to communicate and manage the rest of the home appliances and offers to users' tools to reduce their consumption. Using context-aware information in HEMS will provide knowledge of the environment which can be used to further decrease energy consumption.

Section II-A will present the goals HEMS should achieve and the high-level requirements it should fulfil. Section II-B will present the major challenges when designing HEMS.

A. HEMS High-Level Objectives and Requirements

The main objectives of HEMS are shown in Fig. 1. HEMS main goal is to reduce the energy consumption. However, to achieve this, monitoring energy consumption and managing appliances are needed. In order to reduce energy consumption, first it is necessary to know how energy is consumed. Therefore monitoring is needed. Secondly, it is necessary to manage the appliances to apply energy reduction strategies.

We consider that HEMS has to fulfil the requirements summarized in Fig. 2 to achieve these goals satisfactorily:

- Easy to deploy: It has to be taken into consideration that HEMS should be easy to deploy into users' houses because deploying new cables or infrastructure is not the best solution. This requires using already installed communication systems, such as wireless communication or power line communication which will minimize the costs and gain users' acceptance.
- Interoperability: in order to monitor and manage users' appliances efficiently a home network has to be introduced where devices can exchange information and commands without interoperability conflicts.
- Data security: Security has to be incorporated into HEMS in terms of data encryption and authentication to protect the system against external threats. However, security issues will not be analyzed as they are out of scope of this paper.

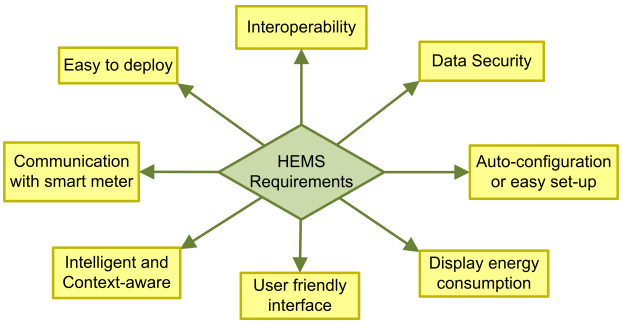


Figure 2. Home Energy Management System Requirements

- Auto-configuration or easy set-up: HEMS is going to be used by users that may not have enough knowledge to perform difficult network configuration tasks. Taking into consideration that users may add or change their home appliances, HEMS should provide easy to use configuration tools or in the best case the network should be auto-configurable.
- Display energy consumption: One of HEMS goals is to monitor energy consumption. This information should be available to users through the user interface.
- User friendly interface: The user interface should provide information about the current consumption and also previous consumptions, providing daily, monthly and even annual reports. Additionally, it can offer the possibility to compare the electricity consumption between months or even compare it to other sources, such as average neighbourhood consumption or other users' consumption. This option could be a new service provided by the smart grid through the smart meter. This interface should also provide management options, where the users can modify their preferences and control their appliances. User preferences are related to the strategy used to reduce users' energy consumption and can vary from system to system. Providing the possibility of controlling devices is also important as the system may apply undesired configurations and the user has to be able to correct them.
- Intelligent and context-aware: HEMS should have some intelligence to facilitate efficient energy management. This can be achieved by creating a context-aware system. A context-aware system is capable of collecting information from the environment, or context, and react accordingly. We consider that a context aware system can significantly improve the reduction of energy con-

 ISBN:978-1-61208-113-7

sumption. There are different ways in which context-aware systems can be implemented in HEMS: by defining energy polices or rules or by creating ubiquitous computing system.

In HEMS, a system that uses energy policies is a context-aware system which collects information from the environment and then uses this information together with the rules to reduce energy consumption.

Ubiquitous computing requires a more complex system. We define that a HEMS using ubiquitous computing reduces energy consumption by using context-aware information to predict users' behaviour and then applies the energy management strategy without compromising the users' comfort. Before being able to predict users' behaviour and apply the energy management strategy, there has to be a learning process. This learning process includes (1) collecting context-aware information, which can include location-aware information, and (2) analyze and process this information to extract the users' routines and patterns. Once the learning process is completed, the system can extract the settings needed to reduce energy consumption.

- Communication with smart meter: Enabling this communication will provide the user with real-time price and billing status, energy consumption information, as well as possible services that may arise. An example of a new service could be comparing the household energy consumption to other users' consumption.

In the next section the challenges found when designing HEMS when trying to comply with the above requirements will presented.

B. Issues and challenges

The main challenge to provide an efficient HEMS is interoperability. HEMS should provide seamless interaction between devices. However, there are a number of different home appliance manufacturers and communication technologies available for the user which makes device interoperability problematic. In addition, devices of the same type, such as washing machines, can have different functionalities depending on the model. Technical incompatibility has limited market possibilities. Users are looking for a 'one size fits all' solution without having to worry about compatibility requirements. An example of how to solve this problem can be found in [4].

Additionally, there are other challenging users' expectations that have to be fulfilled related to the following requirements: auto-configuration or easy set-up, user friendly interface and easy to deploy:

- Easy to use and easy device control: there is diversity in users' preferences and expectations when interacting with HEMS. Some users would like an interface that will give them advanced options while others would just like a simple system but without losing control of their devices [5]. Furthermore, users have different user interface display preferences, some users would like to use their mobile phone or PDA, while others would rather use their computer or a controller. An example of how to deal with this can be found in [6].
- Easy to configure: complex configuration or need of a professional to configure the network is a drawback. HEMS should be easy to configure or even be auto-configurable. However this can be a challenge due to the heterogeneity of home appliances and home technologies.
- Easy software upgrade: home appliances can have software installed, which in some occasions has to be updated. Software update should be easy for users to do. An example of how to deal with this can be found in [7].

Moreover, designing HEMS as an intelligent and context-aware system is not an easy task and presents the following challenges:

- Design of context-aware systems, data collection and interpretation: HEMS may use sensors to collect information about the users' behaviour. The system may have to work with different types of sensors and from different brands. This will force the system to be designed to deal with different sensor details which sets a barrier to interoperability. [8] proposes an infrastructure to support software design and execution of context-aware applications using sensors to collect data.

  Another issue is coping with the amount of data transmitted from home appliances and possible sensors. An example of how to deal with this can be found in [9].
- Policies and rules: There are two main challenges when using policies to implement energy management: coordination and contradiction. As the number of appliances in the house increases so does the number of policies, which can lead to coordination problems and contradictory rules. Tools to identify interactions and detect contradiction between policies should be incorporated into HEMS to manage rules and policies more efficiently. An example of this can be found in [10] and [11].
- Ubiquitous computing: HEMS using ubiquitous computing should include an algorithm which after processing the collected data will be able to learn and predict the users' behaviour . Examples of such algorithms can be found in [12] and [13].
- Multiple-inhabitants: Prediction of users' behaviour when there is more than one user in the home environment adds complexity to the predicting algorithm as each user has his/her own routines and practices.
- Not compromising users' comfort: HEMS should not have undesirable outcomes, it should be an intelligent system that can adapt to different situations and user behaviour.

 ISBN:978-1-61208-113-7

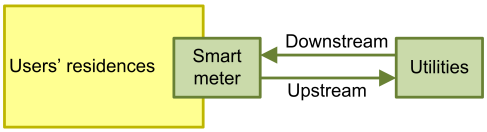


Figure 3. Communication between users' residences and utilities

III. Consumers and Utilities cooperation

Efficient distribution and production of energy requires knowledge about users' energy demand. Collecting data of the energy consumed in households can help built up statistical information about consumption patterns. By providing this knowledge to the utilities, they can foresee the energy needs of their consumers and avoid electricity shortages or blackouts.

Deployment of renewable energies, such as photovoltaic panels or solar thermal panels, is increasing in home environment. These renewable energies help reduce the user electricity consumption from the grid, and, if excess of electricity is produced, they can insert electricity into the grid. Utilities can also use renewable energy production data to foresee the energy demand.

Furthermore, it is expected that electric vehicles will replace fuel vehicles. These new electric vehicles are going to be equipped with a battery that will be charged directly from the grid. When demand peaks occur the users can use the electricity already stored in their vehicles to lower the load in the grid.

In order to collect all these data a cooperation between consumers and utilities has to be established. This cooperation involves incorporating a bidirectional communication system between utilities and users. Users will also benefit from this communication as utilities can provide them real time information about electricity price and billing status. In addition, there will be no need for having utilities employers coming to read the electrical meter as that information will be obtained through this communication system providing Advance Metering Reading (AMR).

As shown in Fig. 3, the upstream communication is define as the transmission of data from the user to the provider and the downstream is defined as the one from provider to user. As stated before, the data transmitted in the upstream will include information about users' electricity patterns taking into consideration their renewable energies. The downstream communication is the transmission of electricity price and billing information from the utilities to the users. Having access to real time price and billing information will make the users become more conscious about their electricity consumption and they may try to reduce the associated costs, by avoiding peak hours, leading to a more distributed and efficient consumption. In addition, the utilities can use this downstream to ask their users to reduce their demand when demand peaks occur. This communication system will enable utilities to be proactive, acting before the problem occurs instead of reacting to it. Furthermore, utilities can offer new service that can be access by the user through this downstream.

This bidirectional communication channel will have benefits for both sides.

A. Requirements

To enable a bidirectional communication between users' and utilities some changes have to be performed in users' residences and in the electrical grid. Traditional electricity meters have to be upgraded to make communication between users and utilities possible. Smart meters are similar to traditional meters, they collect data about the users' consumption, but they additionally support communication between utilities and users. In addition to smart meters, an Advanced Metering Infrastructure (AMI) has to be deployed in the electricity grid. AMI is a system capable of measuring, collecting and analyzing energy usage which is expected to be deployed within the electrical grid. Smart meters comprise a major component of AMI and one of the first necessary steps for bidirectional communication between users and utilities.

B. Issues

AMI is a communication infrastructure that can involve the communication of different utility sectors and companies. The electrical grid is mainly divided into: (1) generation: production of electricity, (2) transmission: transmission of electricity from generators to distribution systems, (3) distribution: connection of power lines to consumers, and (4) consumers. Deploying AMI will require that these parties work together to obtain the maximum benefits. This may require data interfaces between the different parties to deal with interoperability issues. Integrating AMI into the grid may require (1) to deploy a new communication infrastructure in the grid, (2) to use wireless networks such as mobile networks, (3) to use web-based communication. In addition, the communication should be secure to prevent cyber-attacks.

IV. Energy Management in Electric Grids

As stated before, energy consumption in home environments is increasing and consumption patterns have considerably changed in the last years. However, the electrical grids have not changed significantly during the last century, therefore an upgrade is needed to achieve efficient energy distribution and production. This upgrade in the electrical grid will lead to the so called smart grid. [14] defines the smart grid as "electricity networks that can intelligently integrate the behaviour and actions of all users connected to it - generators, consumers and those that do both - in order to efficiently deliver sustainable, economic and secure electricity supplies". Smart grids will incorporate AMI and agents to fulfil the requirements for energy efficiency. In the next section, the objectives of the smart grid are described.

 ISBN:978-1-61208-113-7

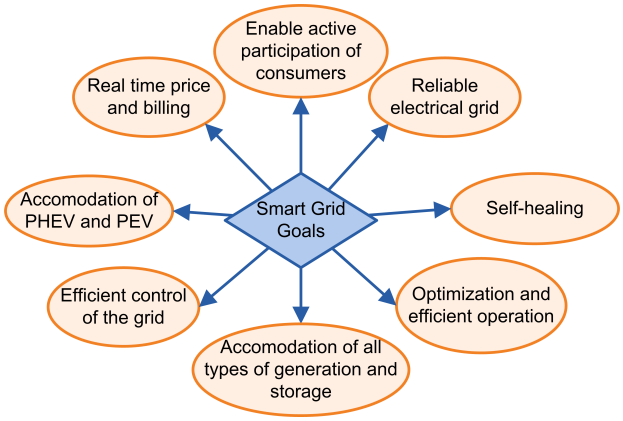


Figure 4. Smart Grid Goals

A. Smart Grid Objectives

Fig. 4 summarizes the main high-level objectives the smart grid should fulfil. When designing the smart grid these goals should be taken into consideration and be integrated together in order to maximize the benefits.

- Enable active participation of consumers: The grid should accept injections from renewable energies installed in the users' residences. In addition, the grid can ask the users to reduce their consumption to avoid demand peaks or shortages and reward them with economical benefits. This process is referred as Demand Response (DR).
- Reliable electrical grid: the smart grid should improve security and quality of supply and reduce the number of blackouts and shortages.
- Self-healing: the smart grid should be more resilient than the electrical grid we have today. Smart grids should be easily reconfigurable and dynamic in order to achieve self-healing which will make the grid more resilient and reliable.
- Optimization and efficient operation: Optimization and efficient operation of the grid implies a reduction of energy losses in power lines. This can be achieved by upgrading the grid components and by using consumption statistics to foresee the electricity usage.
- Accommodation of all types of generation and storage: The smart grid has to accommodate from large centralized power plants to renewable energies installed in the users' premises or distribution systems. In addition, it is foreseen that new storage systems, such as community storages, may be included in the smart grid. To properly manage and control these new elements, the smart grid should be designed as a decentralized and distributed grid.
- Efficient control of the grid: As explained in the previous section, AMI will enable the transmission of energy consumption which can be used by the utilities to forecast users' demands. This information will enable an efficient control of the grid as load scheduling and management can be carried out to avoid blackouts and electricity shortages. In addition, smart grids should enable control information communication among different elements of the grid to achieve an efficient control of the grid.
- Accommodation of PHEV and PEV: Even though Plug-in Hybrid Electric Vehicles (PHEVs) and Plug-in Electric Vehicles (PEVs) are not yet wide-scale adopted, they should be taken into consideration when designing the smart grid. It is foreseen that the amount of PHEV and PEV will increase which will lead to a considerable increase of electricity demand.
- Real time price and billing: AMI will provide the infrastructure to transmit real time price and billing information to the user. The smart grid should incorporate the necessary elements to make this information available such as billing databases.

B. Towards Smart Grid

The electrical grid has to undertake a transformation to reach the smart grids objectives. Introducing AMI into the grid will provide the communication tools to help reach some of the smart grids objectives. However, further changes in the smart grid components have to be done to successfully fulfil these goals. Advanced components, advanced control methods and communication and improved decision support will be introduced in the electrical grid as it moves towards becoming a smart grid. In addition, sensing and measurements technologies should also be incorporated to evaluate the correct functionality of all elements in the grid and enable and efficient control.

V. Conclusion

There is considerable literature on energy management and smart grid. This paper has tried to outline the main goals that have to be fulfilled by the Home Energy Management System, AMI and smart grids. When developing systems to reduce or make energy consumption more efficient, such systems usually focus on one specific capability. It is important that the overall framework and objectives are taken into consideration during the design of such systems to maximize their benefits. This paper can be upstream communication theused as a guideline of the objectives that should be fulfilled by HEMS and smart grids.

References

[1] B. Consortium, "D2.1: Service requirement specification," 2009.

[2] E. Commission, *European Technology Platform SmartGrids - Vision and Strategy for Europe's Electricity Networks of the Future*. Office for Official Publications of the European Communities, 2006.

[3] H. Kudo, "Energy conservation technologies and expectation in japan," 2008.

 ISBN:978-1-61208-113-7

[4] D. Bonino, E. Castellina, and F. Corno, "The dog gateway: enabling ontology-based intelligent domotic environments," *Consumer Electronics, IEEE Transactions on*, vol. 54, no. 4, pp. 1656 –1664, 2008.

[5] L. T. McCalley, C. J. H. Midden, and K. Haagdorens, "Computing systems for household energy conservation: Consumer response and social ecological considerations," in *in Proceedings of CHI 2005 Workshop on Social Implications of Ubiquitous Computing*, 2005.

[6] R. Kistler, S. Knauth, D. Kaslin, and A. Klapproth, "Caruso - towards a context-sensitive architecture for unified supervision and control," in *Emerging Technologies and Factory Automation, 2007. ETFA. IEEE Conference on*, 25-28 2007, pp. 1445 –1448.

[7] S. Grilli, A. Villa, and C. Kavadias, "Comanche: An architecture for software configuration management in the home environment," in *NBiS '08: Proceedings of the 2nd international conference on Network-Based Information Systems*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 283–292.

[8] G. D. A. Anind K. Dey, Daniel Salber, "A context-based infrastructure for smart environments," Georgia Institute of Technology, Tech. Rep., 1999. [Online]. Available: http://hdl.handle.net/1853/3406

[9] N. Shah, C.-F. Tsai, and K.-M. Chao, "Monitoring appliances sensor data in home environment: Issues and challanges," in *Commerce and Enterprise Computing, 2009. CEC '09. IEEE Conference on*, 20-23 2009, pp. 439 –444.

[10] M. Shehata, A. Ebertein, and A. Fapojuwo, "Iris-ts: Detecting interactions between requirements in doors." *INFOCOMP Journal of Computer Science*, vol. 5, no. 4, pp. 34–43, 2006.

[11] M. Shehata, A. Eberlein, and A. Fapojuwo, "Managing policy interactions in knx-based smart homes," vol. 2, pp. 367 –378, 2007.

[12] M. M. Hua Si, Shunsuke Saruwatari and H. Morikawa, "A ubiquitous power management system to balance energy saving and response time based on device-level usage prediction," *IPSJ Journal*, vol. 18, pp. 147–163, 2010.

[13] S. K. Das and D. J. Cook, "Designing smart environments: A paradigm based on learning and prediction," in *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, 2005.

[14] E. T. P. S. for the Electricity networks of the Future. [Online]. Available: http://www.smartgrids.eu/?q=node/163, accesed June 2010

 ISBN:978-1-61208-113-7

Experimental Assessment of Routing for Grid and Cloud

Douglas O. Balen, Carlos B. Westphall, and Carla M. Westphall
Network and Management Laboratory
Federal University of Santa Catarina
Florianópolis – SC - Brazil
Emails: {douglasb, westphal, carlamw}@inf.ufsc.br

Abstract— **Grid and Cloud computing technologies are being applied as an affordable method to cluster computational power together. These structures aim to support service applications by grouping devices and shared resources in one large computational unit. However, the management complexity grows proportionally to the number of resources being integrated. The paper claims to address the problems of management, considering the routing problem in a particular context. An experimental assessment of routing for grid and cloud is presented. In addition, it introduces a proof-of-concept implementation and case study scenarios.**

Keywords - Grid and cloud computing; autonomic systems; routing; network management.

I. INTRODUCTION

Since the creation of the Internet, systems have become increasingly complex due to the scalability and availability requirements posed by several of today's Web services. The popularity of pervasive computing also contributes to increase this complexity as new portable devices are routinely released in the market and integrated into the Internet Cloud.

According to IBM [1], traditionally, networks and management systems are manually controlled processes which demand one or more human operators to manage all the computing systems aspects. In this environment, the operator is strongly integrated to the management process and his task is to execute low level system calls to solve imminent problems. Even though this kind of management, which keeps a human into the system, was appropriate in the past, it cannot cope with modern systems.

The need to connect many heterogeneous systems is one of the main necessities of grid and cloud computing, introducing new levels of complexity. Even though it is a complex environment, the configuration and management is done by humans. This characteristic makes this task slow and a subject of decision making problems. Even administrator errors can occur at this task. In order to avoid this problem a solution is needed in which the management does not need human intervention. Observing this scenario, a question emerges: How to manage efficiently and in an automated way a heterogeneous and complex environment, like grid or cloud?

In order to answer this question this work proposes an experimental assessment of routing for grid and cloud computing that supports autonomic computing paradigm. The system has self-management properties, and redefines the human operator's responsibilities, where their experience is used to define general objectives and polices to control the system instead of placing them in a decision making position.

The rest of this work is organized as follows: Section II provides some comments on autonomic computing, and Section III discusses grid and cloud computing. Section IV proposes an experimental assessment of routing for grid and cloud computing and Section V describes the implementation and tests performed.

II. AUTONOMIC COMPUTING

An autonomic system is able to regulate its own functional parameters without incurring changes in the main system objectives. This way an autonomic system can optimize the use of its resources even under stress conditions. As described by Horn [2], to achieve complete autonomy a system must implement four main characteristics: self-configuration, self-healing, self-optimization, and self-protection.

The autonomic elements (AE), considered to be like the bricks of a building, are the functional units of autonomic systems. They control the resources and offer services to the users and other AEs. They also manage the internal behavior and its relations with other elements of the system, like the policies established by humans or other AEs. The autonomic behavior of the whole system emerges from the numerous interactions between the autonomic elements. An autonomic element consists of one or more managed elements, linked to a single autonomic manager (AM) that controls the managed elements, as shown in Figure 1. The managed elements can be a hardware or software resource.

What differentiates an autonomic from a non-autonomic system is the presence of the autonomic manager. Between monitoring of managed elements and its external environment, the autonomic manager is able to build and execute plans based on the analysis of sent information, which removes the need for human intervention.

 ISBN:978-1-61208-113-7

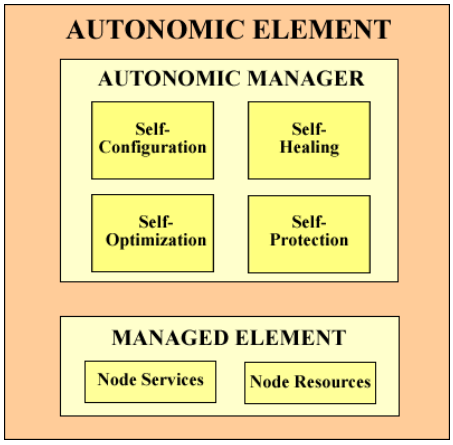


Figure 1. Autonomic Computing Element

III. GRID AND CLOUD COMPUTING

Grid and cloud computing solutions aim to simplify the access to resources (hardware and software) of a distributed system, some times giving the idea that they form a unique and powerful computer. This is achieved by techniques such as virtualization. Resource virtualization [3] minimizes the impact of heterogeneity by providing access to well defined interfaces or to work units in terms of virtual machines. Using this set of abstractions the user can connect several different devices on his network.

The middleware is the software layer between the operational system and the applications, which provides some services that are needed by the applications. It creates the grid environment and gives transparency to the applications. There are several projects in this field, including Globus [4], Gridbus [5], Legion [6], UNICORE [7], Alchemi [8], OurGrid [9] and Grid-M [10].

TABLE I
SOME CURRENT MIDDLEWARES

	Collaboration Support	Context Awareness Support	Resource Allocation Support	Dinamic Enviroument Support	Mobile Enviroument Support	Self-Management Support
Globus	■		■			
Legion	■		■			
Gridbus	■		■		■	
UNICORE	■		■			
Alchemi	■		■			
OurGrid	■		■			
Grid-M	■	■	■	■	■	

Looking at the features supported by these systems in Table 1, one can see that all listed middleware support collaboration and resource allocation. However, only two systems support execution on mobile environments. Only one provides context sensibility. None of them supports autonomic behavior. Due to a grid complexity, there is a need for middleware that supports autonomic.

There is no consensus about what exactly cloud computing is, but some characteristics are clearly repeated. It is a new distributed computing and business paradigm. It can provide computing power, software and storage resources, and even a distributed data center infrastructure on demand. To make these characteristics viable, it uses existing technologies, such as virtualization, distributed computing, grid computing, utility computing and the network infrastructure provided by the Internet. In this work, we are considering cloud computing, using our middleware Grid-M [10], developed by the Laboratory on Networks and Management.

IV. ROUTING FOR GRID AND CLOUD COMPUTING

Pervasive computing is a paradigm that aims to provide a computing environment anywhere through the use of virtualization of information, services and applications.

A middleware capable of supporting this new computational environment must offer large scale distributed computing that permits to integrate sensors and mobile devices, always taking into consideration the dynamics of the environment and the context sensibility.

The only middleware from those examined that presents these characteristics is the Grid-M [10]. However, similarly to others, it does not offer autonomic behavior. The computational grids are known as a dynamic and heterogeneous computational environment, even though, the configuration of these environments is done manually and susceptible to slow decision making or errors of the administrators. In order to avoid this problem a solution is needed to take the responsibility away from the human administrators.

This work proposes a system for this kind of environment, offering the opportunity to create a grid and cloud computing with autonomic management.

A. *Related Work*

The system proposed intersects with fields that are being the target of continuous academic research such as autonomic systems, and grid and cloud computing. However, the union of these initiatives is still new and related work with the same focus is scarce. Some of the projects in this area are:

- Liu et al. [11] proposes an autonomic architecture to manage the heterogeneity and dynamics of the grid environments. This architecture allows the behavior of services and applications and its interactions to be specified and adapted according to the high-level rules. Everything is based on the requisites, states and execution context of the applications;
- Beckstein et al. [12] presents the SOGOS architecture aimed to support self-organization in computational grids. This is allowed to work with dynamic environments through semantic information

 ISBN:978-1-61208-113-7

(metadata) that describes the involved organizations, roles, rights of the participating agents and how they interact to solve the problem. The decisions are based on the metadata;

- Brennand et al. [13] presents the AutoMan, a system which has the objective of offering certain levels of automatic management to the computational grids in pairs. Beyond this scope, it tries to optimize the usage of resources on the grid, simplifying the management activities at the same time;
- Buyya et al. [14] defines Cloud computing and provides the architecture for creating market-oriented Clouds by leveraging technologies such as Virtual Machines (VMs);
- Xiao et al. [15] adapts web pages to small screen devices. In addition, as the limited computing ability and capacity of storage of wireless handheld devices, it is extremely challenging to deploy existing web page adaptation engine. By utilizing the large computing and storage resource capabilities of cloud computing infrastructures, a new wireless web access mode is proposed; and
- Vieira et al. [16] shows a solution for intrusion detection in grid and cloud computing environment in which audit data is collected from the cloud and two intrusion detection techniques are applied.

B. *Autonomic Manager*

What allows a system to be called autonomic is a presence of an autonomic manager. Through the monitoring of managed elements and their external environment, the autonomic manager is able to build and execute plans for implementation, based on the analysis of sent information. Therefore, the autonomic manager is responsible for ensuring self-management, achieved when all its sub-areas (self-configuration, self-regeneration, self-optimization and self-protection) are guaranteed.

For this purpose, this paper suggests that the manager is composed of some components, responsible for monitoring the data sent by the managed elements and others elements of the autonomic grid, analyze them, plan actions according to their objectives and implement these actions, thus achieving a high degree of autonomy.

C. *Routing among the nodes*

The number of mobile devices is constantly changing, which can result in big changes in the overall system. For the interconnection among the devices, it is essential to keep the routing table consistent. The Routing Table Management component has the goal of detecting routing inconsistencies, but it cannot directly manipulate the routing table. The latter is done by the grid's routing algorithm.

The system proposed here implements two routing algorithms: one is based on the direct interconnection with a neighbor node, and the other is based on the interconnection among all nodes.

In grids, every element has its own routing table that contains the destination (node name) and a metric (the distance until the next element in hops). On the first algorithm, each node connects to the neighbor node only. Thus, the route to the neighbor node becomes a default route (gateway) to the other elements in the grid. For example, when an element wants to request a service, it sends a request to the gateway, and the gateway is responsible for forwarding the request to the others nodes connected to it. This process is repeated until the destination receives the request.

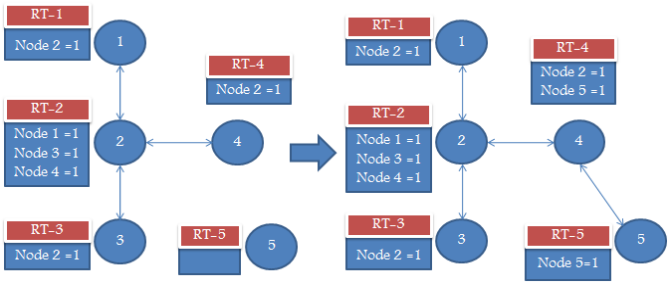


Figure 2. Routing algorithm based on the direct interconnection to the neighbor node.

Figure 2 illustrates how that algorithm works. On the left side, node 5 is out of the grid, thus the other elements cannot connect to it. As soon as node 5 joins the grid through node 4, the latter includes a route to node 5 with metric 1, i.e., directly connected.

The other algorithm is a little different. As an element joins the grid, all the other elements add a direct route to it (metric 1). This makes the whole grid to be seen as a complete graph. The propagation of the information about a node joining or leaving the grid is coordinated by this same algorithm in an autonomic way. When all the nodes discover the topology changes, we have reached the convergence.

Figure 3 illustrates this situation. At first, node 5 is out of the grid. Note that all other elements are directly connected (metric 1). Then, node 5 is included. It does not matter knowing which node it is connected to, because the distance among all elements is the same. The first node to notice its join-request is going to add a direct route to it, sends its actual routing table, and finally informs all the other elements that there is a new node in the grid.

 ISBN:978-1-61208-113-7

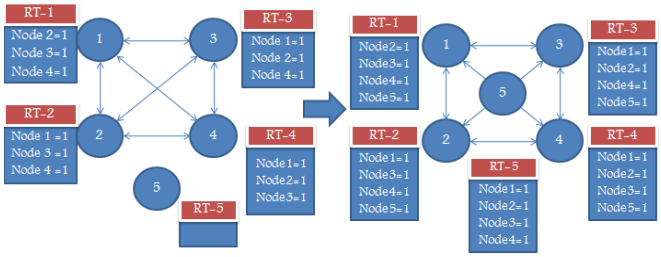


Figure 3. Routing algorithm based on the complete connection among nodes

It is the responsibility of the grid's manager to decide which algorithm to use. Remark that it is not possible to use both algorithms at once, all nodes must use the same algorithm.

V. IMPLEMENTATION AND TESTS

To this point, this paper described the theory upon which the proposed system was based, the architecture details, its components and interactions, and the routing algorithms. To test it, we have implemented it on Grid-M [10]. Among the main benefits of the Grid-M middleware are: it is open source, it is easy to deal with small devices, it has a friendly API and it is portable [10].

This section shows the results of a few quantity tests performed during the implementation with the purpose of showing the proposed system efficiency in different use situations.

A grid of 30 nodes was created. These devices are personal computers with an Intel Core Duo 1.66Ghz CPU, 2GB of RAM memory and running Window XP. All devices ran the same programs.

A. *Convergence Time*

Here, we do three separated tests for the two kinds of algorithms to test the convergence time. To a routing protocol, convergence time means the time it takes for all the routing tables to be updated when there is a change on the topology (e.g., when a node joins the grid).

At the beginning, we thought the convergence time would be a bottleneck, especially on the algorithm which all nodes are directly connected since all routing tables are spread among all nodes.

Analyzing Figure 4 though, which shows the convergence time of the algorithm based on the direct interconnection to the neighbor node, we notice that the convergence time is really small and almost constant (varying between 10ms and 14ms). This happens because the only processing needed is the inclusion of the neighbor's route in the routing table. No data about a joining node is passed along. The time was taken when a new element joined the grid. The elements were added in the following manner: node 2 connects to node 1, node 3 connects to node 2, node 4 connects to node 3, and successively.

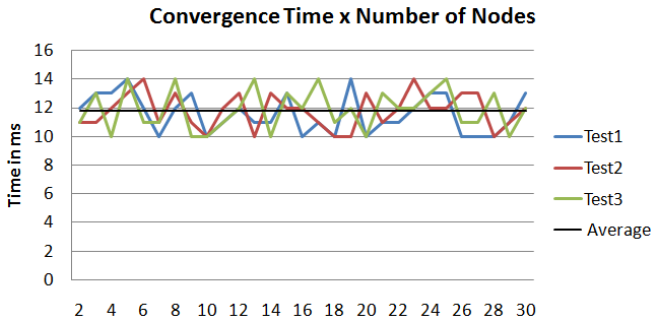


Figure 4. Convergence time – Algorithm based on the direct interconnection to the neighbor node

On the other hand, on the other algorithm, when a node joins the grid, all elements' routing tables are updated with the new information. The convergence time of this algorithm is shown on Figure 5. The data was obtained the same way as the previous test.

Figure 5 shows that the lowest convergence time was achieved on test 2, after the insertion of node 6, and the highest convergence time was achieved on test 2 as well, after the insertion of node 10. As you can see, as more nodes get in the grid, the convergence time increases, but on an ease pace (the average time at the beginning was 138ms and at the end it was 144ms). As the convergence time was still low in this case, we chose this algorithm because its response time during tasks executions is a lot lower.

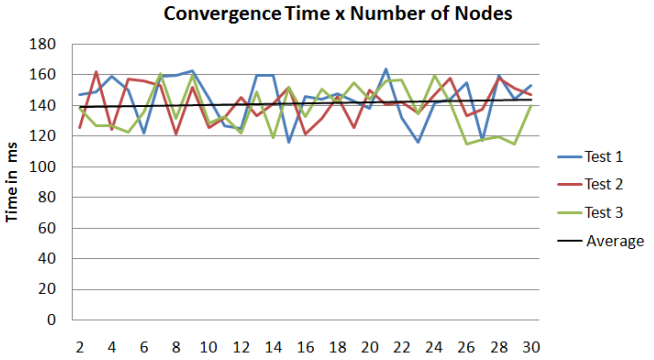


Figure 5. Convergence time – Algorithm based on the complete connection among the nodes

B. *Response time of the tasks execution*

Another important point is the response time of a service request. The response time refers to the sum of two distinct tasks: the service search time and the task execution time.

The task search time is the time a requester spent to search for a determined service in the grid to find who has it and who has the best available resource percentage. So, at the end of the search, we have the best candidate for the execution and a task is created with him being the destination. All the process of search and request redirections is managed and controlled by the Autonomic Manager.

With the intention of testing the response time of the service execution requests, we have used the same structure of the previous test (30 equal nodes). The test consists in the node 1 request a service to the grid. The only node that has it is the node 30.

We would like to clarify that the response time depends on the routing algorithm type utilized. As we use the algorithm based on the direct interconnection with the neighbor node, the search takes longer if we compare it to the algorithm based on the complete interconnection among nodes. This happens because the latter has a complete view of the topology. Therefore, the search in all nodes can be done in parallel (by using threads). Case we use the first algorithm, the search request must pass through the intermediate nodes before getting to its destination. The test results using both algorithms are shown on Figure 6.

As expected, the response time of the algorithm based on the restrict connection to the neighbor node is longer than the other one.

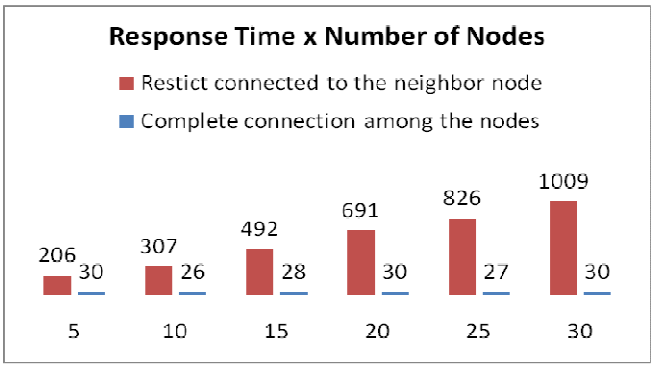


Figure 6. Response time results

C. *Efficiency of the services replication*

On the previous test, the node 1 requested a service to the grid. When the search was done, it was verified that only node 30 offered the determined service. As it was a test and we knew that there was only one requester, we discard the possibility of the node 30 being a bottleneck due to it being overloaded. However, what would happen if the other 29 nodes requested the same service? On this case, there would be the possibility of the node 30 not being able to answer to all requests on the best possible way, lowering the performance of the grid. At this time the node 30, aware that he is overloaded, would send a replication request to find an available element that offers the same service. Note that the replication is necessary only once. After that, the node that received the service will start answering to other requests about the same service.

Figure 7 shows the resources used by 4 elements during the tests. We got this information from the Grid-M logs.

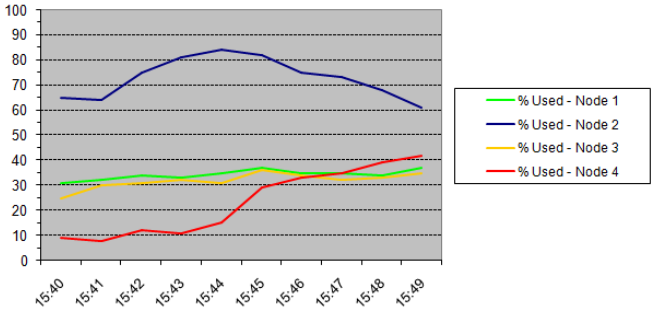


Figure 7. The resources utilized by the nodes and the services replications

On this test, nodes 1 and 3 make requests all the time to a service that initially only node 2 provides. After a while, node 2 becomes overloaded (the free resources percentage gets lower than 18%) and then a service replication occurs, from node 2 that has the service, to node 4 that was the node which had more free resources at that time. After that, the service requests are answered by node 4 as well, distributing the processing of these requests. Analyzing the chart (Figure 7), we observe that the algorithm eliminated the eminent saturation of the node 2 and the possible creation of a bottleneck in the grid.

VI. CONCLUSION

In this paper, we have proposed an experimental assessment of routing for grid and cloud computing. The convergence time of the algorithm based on the direct interconnection to the neighbor node is really small and almost constant. As expected, the response time of the algorithm based on the restrict connection to the neighbor node is longer than the other one. The big question to be answered was: How to make a heterogeneous environment and with huge complexity, like grid and cloud computing, not being managed manually, which is inefficient? The solution proposal is the creation of autonomic elements acting as intelligent agents, capable of feel the environment where they are and act the same according to pre-defined policies.

 ISBN:978-1-61208-113-7

REFERENCES

[1] IBM-Corporation. An architectural blueprint for autonomic computing. *http://www.ibm.com/developerworks/autonomic/library/ac-summary/ac-blue.html*

[2] P. Horn. Autonomic computing: IBM's perspective on the state of information technology. Technical report, *International Business Machines Corporation*, Armonk, NY, USA, 2001.

[3] J. Joseph and M. Ernest. Evolution of grid computing architecture and grid adoption models. *IBM Systems Journal*, 2004, 43(4).

[4] I. Foster and C. Kesselman. Globus: A metacomputing infrastructure toolkit. *Internacional Journal of Supercomputer Applications*, 1997, 11(2), pp. 115–128.

[5] R. Buyya. Market-oriented grid computing and the gridbus middleware. *16th International Conference on Advanced Computing and Communications*, 2008. ADCOM 2008.

[6] A. Grimshaw and A. Natrajan. Legion: Lessons learned building a grid operating system. *Proceedings of the IEEE*, 2005, 93(3), pp. 589–603.

[7] UNICORE. UNIform Interface to Computer Resources. http://www.unicore.eu/ (last access on Dec. 2010).

[8] A. Luther, R. Buyya, R. Ranjan, and S. Venugopal. Alchemi: A .net-based grid computing framework and its integration into global grids, pp. 1-17, 2005. (informal publication). http://www.cloudbus.org/papers/Alchemi.pdf

[9] F. Brasileiro, E. C. de Araujo, W. Voorsluys, M. Oliveira, and F. Figueiredo, "Bridging the High Performance Computing Gap: the OurGrid Experience," ccgrid, pp.817-822, Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid '07), 2007.

[10] H. A. Franke, C. Rolim, C. B. Westphall, F. Koch, and D. O. Balen. Grid-M: Middleware to integrate mobile devices, sensors and grid computing. *The Third International Conference on Wireless and Mobile Comunications – ICWMC* 2007.

[11] H. Liu,V. Bhat, M. Parashar, and S. Klasky. An autonomic service architecture for self-managing grid applications. InGRID'05: *Proceedings of the 6th IEEE/ACM Internation Workshop on Grid Computing*, 2005.

[12] C. Beckstein, P. Dittrich, C. Erfurth, D. Fey, B. Konig-Ries, M. Mundhenk, and H. Sack. Sogos-distributed meta level architecture for the self-organizing grid of services. *In MDM'06: Proceedings of the 7th International Conference on Mobile Data Management*, 2006.

[13] C. Brennand, M. Spohn, A. Souza, G. Ferreira, D. Candeia, G. Germoglio, and F. Santos. Automan: Autonomic Management on Ourgrid. *V Workshop for Grid Computing and Aplications,* 2007.

[14] R. Buyya, C. S. Yeo, and S. Venugopal, Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities. *10th IEEE International Conference In High Performance Computing and Communications*, 2008.

[15] Y. Xiao, Y. Tao, and Q. Li, A New Wireless Web Access Mode Based on Cloud Computing. *Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, 2008.

[16] K. Vieira, A. Schulter, C. B. Westphall, and C. M. Westphall, "Intrusion Detection for Grid and Cloud Computing". IEEE IT Professional Magazine. V.12 (4). pp. 38-43. 2010.

 ISBN:978-1-61208-113-7

Scalability of Distributed Dynamic Load Balancing Mechanisms

Alcides Calsavara and Luiz A. P. Lima Jr.
Graduate Program on Computer Science (PPGIa)
Pontifical Catholic University of Paraná - PUCPR
Curitiba, Brazil
{alcides, laplima}@ppgia.pucpr.br

***Abstract*—A load balancing mechanism for large-scale systems should be distributed and dynamic in order to accomplish scalability and high availability. Also, it should be autonomic in order to ease network management. The recent development in utility computing architectures, such as the so-called cloud computing platforms, has increased the demand for such mechanisms. This paper investigates a novel approach, based on the concept of virtual magnetic fields, by which ready-to-start tasks launched on a network middleware are "attracted" to idle nodes. The key issue on such approach is that the update of workload information amongst the cooperating nodes of a network must be a low-cost and an autonomic operation. Two different update algorithms are presented, and their complexity is assessed through simulation. The results show that both algorithms fulfill the scalability requirement.**

Keywords - load balancing; scalability; distributed algorithm; autonomic systems

I. INTRODUCTION

The recent development in utility computing architectures, such as the so-called cloud computing platforms, has increased the demand for load balancing mechanisms, which provide for scalability, high availability and ease of management, amongst other requirements [6][7]. Typically, utility computing architectures are deployed as large-scale complex systems, where changes in resource availability happen very often and in an unpredictable way because, in principle, a task launched at some client node can be assigned to any server node, at any time. Moreover, the geographically distributed nature of such systems makes centralized assignment of tasks to specific servers infeasible, as the corresponding scheduler would become a network bottleneck and a single point of failure. Therefore, a key issue in the development of utility computing architectures is the provision of a distributed load balancing mechanism which is able: firstly, to respond within an approximately constant time regardless network size and topology, i.e., it should scale well; secondly, to proceed responding in the presence of failures such as node crash and network partition, i.e., it should be highly available; and, thirdly, to manage workload information change with a minimum of, preferably none, human intervention, i.e., it should be autonomic.

Several distributed load balancing mechanisms have been previously reported in the literature, as discussed in Section II, and, to the best of our knowledge, their suitability for utility computing architectures are yet to be verified. This paper investigates a novel approach, based on the concept of virtual magnetic fields, by which ready-to-start tasks launched on a network middleware are "attracted" to idle nodes.

The paper is organized as follows. Section II discusses some related work and their weaknesses and strengths. A generic formal model for Virtual Magnetic Fields is presented in Section III, and two workload update algorithms (namely, *QuickPath* and *ShortPath*) are described in Section IV, each taking a different approach to the problem. The algorithms are then evaluated and compared through simulation, and the corresponding results are presented in Section V. Finally, Section VI draws some conclusions and discusses future work.

II. RELATED WORK

The problem of load balancing in an open environment, such as a P2P overlay network, is well discussed in [1], where many complex related issues are listed. Amongst them, the problem of resource discovery, which means to search for idle CPU cycles in this case, is considered as extremely difficult since such resource is perishable, cannot be shared, and is dynamic. Moreover, the set of participating hosts is potentially very large and highly dynamic. The authors compared several methods to solve that problem and they noticed that a hard problem to solve is that large jobs may dominate, resulting in delays for scheduling smaller jobs.

The problem of load balancing is also well studied in the context of grid computing. However, differently from P2P computing, where cycles are obtained from ordinary users in a distributed open environment, in grid computing, cycles are normally obtained from a previously known set of users who agreed to share such resource according to well-defined rules. A corresponding scheduling algorithm enforces such rules and, as well, takes care of proper load balancing. Thus, it can be simpler than a scheduling algorithm for P2P computing. As an example, a dynamic tree-based model to represent a grid architecture in order to manage workload is proposed in [2]. Its main purpose is to improve response time of user's submitted applications by ensuring maximal utilization of available resources through a hierarchical load balancing strategy and associated algorithms based on neighborhood properties. The authors claim to have achieved a reduced communication overhead induced by tasks transferring and flow information. Such solution is based on a group manager who receives, in a periodic way, workload information from

 ISBN:978-1-61208-113-7

each network element.

A discussion on various load balancing algorithms in the context of parallel computing can be found in [8]. The authors also present a list of parameters which can be useful to analyze those algorithms, including *stability*: the cost of workload information transfer versus the benefits of a better overall load balance. According to their research, static algorithms – like round-robin and randomized algorithms – are more stable than dynamic algorithms. However, dynamic algorithms show better overall results than static algorithms when all analysis parameters are considered, especially where fault tolerance is a key requirement.

A comparative study of distributed load balancing algorithms is presented in [7]. The authors discuss the importance of such algorithms in the context of cloud computing, although no particular mapping of the algorithms to real computing platforms is presented. Three distinct algorithms, namely Honeybee Foraging, Biased Random Sampling and Active Clustering, are assessed through simulation in an ideal scenario for cloud computing where nodes are typed to accomplish for node heterogeneity, by using task throughput as the main evaluation parameter. According to their experiments, Honeybee Foraging has a much better throughput than Biased Random Walk and Active Clustering when the number of node types increases in a fixed-size number of nodes; the throughput is slightly better for Active Clustering than for Biased Random Walk. On the other hand, when the number of nodes increases for a fixed-size number of node types, Biased Random Walk and Active Clustering present a much better throughput than Honeybee Foraging, with a small advantage for Biased Random Walk.

III. PROPOSED LOAD BALANCING MECHANISM

Given any two linked nodes, say node *S* and node *T*, a relationship between them can be defined in which *S* attracts messages (which contain tasks) that were initially sent to *T*. Metaphorically speaking, a node S that can attract messages from a neighbor node *T* contains a *virtual magnet* and *T* is within the *virtual magnetic field* produced by such magnet. So, *S* and *T* have a *magnetization relationship*: a source of attraction *S* magnetizes a target *T*. In other words, a magnetization relationship from S to T implies that S can help T to perform its tasks. In particular, any node that possesses a magnet will attract messages to itself. Moreover, the magnetization relationship is defined as transitive, so that, if node *x* (directly) magnetizes node *y* and node *y* (directly) magnetizes node *z*, then node *x* (indirectly) magnetizes node *z*. As a result, the set of all nodes and their magnetization relationships define an overlay network, named as *magnetization network*. For the sake of simplicity, a magnetization network is modeled as a directed graph where each vertex corresponds to a network node and each edge corresponds to a magnetization relationship. The set of nodes which are directly magnetized by a node *x*, denoted as $T(x)$, is discovered by following the edges that leave *x*. Conversely, the set of nodes that directly magnetize a node *x*, denoted as $S(x)$, is discovered by following the edges that enter *x*. By definition, each node is self-magnetized, i.e. any node *x* belongs to both $S(x)$ and $T(x)$. The set of nodes which are either directly or indirectly magnetized by a node *x*, denoted as $T^*(x)$, and, as well, the set of nodes which either directly or indirectly magnetizes a node *x*, denoted as $S^*(x)$, are determined simply by traversing magnetization relationships.

Each node *x* attracts messages according to an associated *strength* or *force*, denoted as $F(x)$, which is set according to local workload information, such as the locally available processing power. So, given a node *x*, the strength $F(x)$ will be determined according to some criteria and, in the general case, it may change over time, as node and network properties change and, as well, as messages are delivered to nodes. For any node *x*, the strongest node in $S^*(x)$ is called the *global pivot* for node *x*, denoted as $P^*(x)$. So, according to the semantics of the magnetization relationship, any message sent to a node *x* must be delivered to $P^*(x)$. Also, for any node *k* in $S(x)$, the *partial pivot for x with respect to k*, denoted as $P_k(x)$, is the global pivot for node *k*, that is, $P_k(x)=P^*(k)$. Thus, given a node *x*, $P^*(x)$ is recursively computed as the node with greatest strength between all nodes $P_k(x)$ where *k* belongs to $S(x)$. Naturally, care must be taken to prevent infinite loops in such computation, since cycles are allowed in the magnetization network. Thus, supposing that tasks are independent from each other and they do not depend on remote data, in the proposed load balancing mechanism, an application message that contains a ready-to-run task is simply attracted, i.e., routed to the node corresponding to the global pivot with respect to the node where the message is originally created. From the implementation point of view, a magnetization network is a network middleware where the life cycle of any application message *m* accomplishes the following steps:

1. *m* is created at the application level and sent to a node *x;*
2. *m* is routed from node *x* to a node *y*, where *y* belongs to $S^*(x)$, such that $F(y)$ is greater or equal to $F(k)$, for any *k* that belongs to $S^*(x)$, that is, $y=P^*(x)$;
3. *m* is delivered to the application level of node *y;*
4. *m* is properly handled at node *y* and, depending on the application semantics, it is either eventually destroyed or kept at node *y* forever.

Every time a change happens with respect to the strength of any node, the magnetic field of such node must be properly updated. Also, as a consequence, the global pivot for any node may change, thus requiring, proper update of its magnetic field, as well. In Section IV, two distinct update algorithms are described and analyzed.

IV. WORKLOAD UPDATE ALGORITHMS

As discussed in Section III, a virtual magnetic field has to be updated when the strength of the corresponding magnet (i.e. its workload) changes significantly, since such magnet strength can be relevant to nodes which are either direct or indirectly magnetized by the node that hosts such magnet. Also, any node can change its global pivot at any time, primarily as a consequence of one or more events of magnet

 ISBN:978-1-61208-113-7

strength change. A global pivot change for a node *x*, on the other hand, can be relevant to the nodes which are directly magnetized by *x*, since their own global pivot can change as a consequence. Hence, this section presents two completely distributed algorithms to keep track of magnetic field information in order to try to guarantee that any application message sent to a node *x* will be attracted to the node that either direct or indirectly magnetizes *x* with the strongest magnet. Naturally, there will always be a chance that an application message is not routed to the ideal (actual strongest) node because the latency of networks may delay the effect of update messages.

Both algorithms work on the basis that there is no centralized component, so every node behaves autonomously and, at the same time, cooperatively with its neighbors with respect to the magnetization network, thus assuming that they can trust each other. That is achieved by defining a common data structure stored by each node in order to keep track of magnetic field information, and a common behavior to internally handle update messages, possibly issuing new ones. Each node is assumed to be uniquely identified and network messages take arbitrary finite time to be delivered, though they never get lost, duplicated or corrupted. In other words, the underlying network protocol does not need to guarantee order, but it has to be reliable.

A. QuickPath

QuickPath is the distributed self-stabilizing algorithm that propagates changes in magnetic fields by updating the strength perception of each node *k* in *T*(x)* using the first notification message that arrives at *k*. The consequence of doing so is that further application messages *m* containing some task that are sent to node *x* will be routed to *P*(x)* through the fastest path (in case there are more than one) determined at magnetic field update time. Once a node *x* receives a notification of magnetic field change, this notification is propagated to *T(x)-{x}* only if either *P*(x)* or *F(P*(x))* have changed. In order to do so, *QuickPath* must avoid remagnetization of nodes. Therefore, if *P*(y)=x*, there must be only one path $k_1, k_2, \ldots, k_n$ from *y* to *x*, in which *P*(* k_i*)=x* ($1 \leq i \leq n$). All other nodes in *T*(x)* that do not belong to $\{k_1, k_2, \ldots, k_n\}$ will be updated with alternative information (typically, second greatest strength and pivot) collected in the update notification path.

This procedure produces two nice properties of *QuickPath*: (a) the algorithm always stabilizes within a finite amount of time for a finite number of nodes (in other words, the number of change notification messages of *QuickPath* is always finite regardless the network topology), and (b) the algorithm updates the perception each node *x* has of *P*(x)* generating acyclic (possibly non-deterministic) routing paths from *x* to *P*(x)*. The first property (a) is consequence of the fact that propagation of change notifications is only carried out if the perceived strength and/or pivot have changed. So, the same message arriving more than once at a node will not cause further propagations. Property (b) derives from the avoidance of remagnetization as explained above.

The core of *QuickPath* can be described as follows. Let *m*={*rm*,*ra*} be a change notification message where *rm*=<*p*,*F(p)*> and *ra*=<*a*,*F(a)*> are pairs with information about the known pivot (*p*) and alternative pivot and their respective strengths (*rm.pivot = p* and *rm.strength=F(p)*). Let *lm* and *la* be similar pairs with information about known pivot and pivot alternative before the receipt of *m*. *QuickPath*'s propagation algorithm is defined by:

```
Node x upon receiving (rm, ra):
begin
   update local info about known pivot and alternative \
      using (rm,ra)
   {lm',la'} = information known after the update
   if (lm'≠lm) or (la'≠la) then
      {pm,pa} = {lm',Max{la',ra}}
      if pm=pa then
         pa = la'      // avoid sending repeated info
      endif
      if pm.strength=pa.strength and pm.pivot>pa.pivot then
         swap(pm,pa)
      endif
      send {pm,pa} to T(x)-{x}
   endif
end
```

The algorithm above always causes propagation of the best alternative pivot known along with de actual pivot information every time the local perception of the magnetic field changes. If the strengths of the main and alternative pivots are the same, always propagate consider pivot the one with smaller *id*, so to break symmetry and avoid propagation loops. Further details of *QuickPath* can be found in [4].

B. ShortPath

In the *ShortPath* algorithm, an application message will traverse the shortest path in case there is more than one path between a node and its global pivot. A more detailed description of the algorithm can be found in [9], where the algorithm is applied for general message routing. The data structure stored by any node *x* consists of the following items:

- *F(x)* : The strength of *x*.
- *S(x)* : The set of nodes that directly magnetizes *x*. For each node *s* in *S(x)*, the following fields are stored: *(1)* The identifier for *s*. *(2)* The identifier for the global pivot for *s*, that is, *P*(s)*. *(3)* The distance from *P*(s)* to *x* with respect to the magnetization network. *(4)* A timestamp corresponding to the local time at *s* when *P*(s)* was set for the last time. Such timestamp is useful to discard related messages which arrive out of order. *(5)* A flag to indicate whether the global pivot information is either up-to-date or obsolete.
- *T(x)* : The set of nodes directly magnetized by *x*. For each node *t* in *T(x)*, the only information stored is the identifier for *t*.
- *K(x)* : The set of nodes known by *x* which either direct or indirectly magnetizes *x*. Hence, *K(x)* is a subset of *S*(x)*. Because of the distributed nature of the algorithm, a node *x* does not know *S*(x)* in advance, so it is discovered as update messages arrive at *x*. For each node *k* in *K(x)*, the following fields are stored: *(1)* The identifier for *k*. *(2)* The strength of *k* currently known by *x*, denoted as $F_x(k)$. *(3)* The timestamp corresponding to the local time at *k* when its strength changed to $F_x(k)$. *(4)* The distance from *k* to *x* with respect to the magnetization network.

 ISBN:978-1-61208-113-7

- *P*(x)* : The global pivot for *x*, which is determined according to the information available on *K(x)*.
- *M(x)* : A node *m* that belongs to *S(x)* such that the strength of the global pivot for *m* is maximum among all global pivots for nodes in *S(x)*. This field is employed to route application messages, since it ultimately leads to *P*(x)*.

All messages exchanged by the algorithm flow according to the magnetization network. So, any message is sent from a given node *x* to a node *y* that is directly magnetized by *x*. There are two types of messages, namely *strength change* and *pivot change*, explained as follows.

A message *m* of type *strength change* contains the following fields: *(1)* The identifier for a sender node *x*. *(2)* The identifier for a destination node *y*. *(3)* The identifier for a node *s* which is the source node whose strength change is being notified by *m*. *(4)* The strength of *s* being notified by *m*, denoted as *F'(s)*. *(5)* A timestamp corresponding to the local time at *s* when its strength changed to *F'(s)*. *(6)* A distance from *s* to *y* with respect to the magnetization network. Such distance is useful for two purposes: (i) to determine the shortest magnetization path between *y* and its global pivot in the case where more than one such path exist; (ii) to detect message loops that may occur due to cycles in the magnetization network. A strength change message *m* which is sent from node *x* to node *y* in order to notify about the strength of a node *s* is handled by node *y* in the following way:

```
if the strength change of s notified by m is relevant
   according to timestamps in m and K(y) then
 1. register into K(y) all data about s contained in m
 2. if the strength of s went down then
      2.1. send a strength change message to every node t
           in T(y) in order to notify about s
      2.2. if s happens to be P*(y) then
          2.2.1. update P*(y) and M(y) according to data on
               S(y)
          2.2.2. if any data regarding P*(y) has changed
               then send a pivot change message to every
               node t in T(y) to notify about P*(y)
```

A message *m* of type *pivot change* contains the following fields: *(1)* The identifier for a sender node *x* which is the source node whose global pivot has changed and is being notified by *m*. *(2)* The identifier for a destination node *y*. *(3)* The identifier for a node*(p* which is the global pivot for *x* being notified by *m*. *(4)* A timestamp corresponding to the local time at *x* when *p* became the global pivot for *x*. *(5) The strength of p* at the time when such node became the global pivot for *x*, denoted as *F'(p)*. *(6)* A timestamp corresponding to the local time at *p* when its strength changed to *F'(p)*. *(7) The distance from p* to *y*. Similarly, to strength change messages, such distance is useful to determine shortest path and to detect message loops. A pivot change message *m* which is sent from node *x* to node *y* in order to notify about the global pivot *p* for node *x* is handled by node *y* in the following way:

```
if the pivot change of x notified by m is relevant
   according to timestamps in m and S(y) then
1. register into S(y) all data about x and p contained in m
2. if either the strength of p notified by m is stale
according to timestamps in m and K(y) or the distance from p
to y notified by m is greater than such distance registered
on K(y)
   then mark x as obsolete on S(y)
   else if the strength of p notified by m is relevant
          according to timestamps in m and K(y) then
           2.1. register into K(y) all data about p
           contained in m
           2.2. if the strength of p went down then send a
           strength change message to every node t in T(y)
           to notify about p
3. update P*(y) and M(y) according to data on S(y)
4. if any data regarding P*(y) has changed then send a pivot
change message to every node t in T(y) to notify about P*(y)
```

V. SIMULATION RESULTS

A preliminary evaluation of the algorithms described in Section IV is presented in this section. The results were obtained from simulation and they show the impact of magnetization network connectivity on the number of update messages, as well as on the time to complete an update cycle and the number of bytes transmitted in each edge. The strength of a node is computed from the amount of resources consumed (memory, CPU, storage space and so on – details are out of the scope of this paper) and it is represented by a value ranging from 0% (the node is completely busy) to 100% (the node is completely idle). For the sake of simplicity, each node has just one server and all servers are capable of processing any task. Simulation was carried out using basically the parameters presented in [3], by randomly scattering a number of nodes over a fixed-sized (1,500x500 meters) rectangular area. The magnetization network was built by assigning a random radial range of influence to each node *x* from 50 to 200 meters; nodes within that range are considered neighbors of *x* (i.e., they are under the magnetic influence of *x*). Next, the magnetization network was a connected graph by consecutively (a) generating its adjacency matrix; (b) computing its transitive closure; (c) generating the set of reachable nodes from each other node; (d) joining sets whose intersection is not empty; (e) joining remaining sets by connecting the nearest nodes belonging to each pair of sets. The number of nodes scattered over the fixed-sized area were multiple of 10, ranging from 10 to 100, in such a way that the magnetization network connectivity and the number of edges per node increased accordingly, thus providing a means to assess its impact on the number of protocol messages. For each number of nodes, 1,000 tests were carried out and the results presented in the sequence correspond to a simple average. Initially, *F(x)* = 100 for all *x* and the measures were taken after producing random strength drops in all nodes simultaneously at instant zero causing a destabilization of *all* magnetic fields.

The number of messages can be large if small strength changes are considered. For that reason, a parameter (not shown in the algorithms), named *threshold* (*th*), was introduced to control when a strength change is relevant, i.e. when a strength change must be notified to the corresponding magnetized nodes. So, when the threshold is set to zero, absolutely any strength change is considered relevant, even the smallest. On the other hand, if, for instance, the threshold is set to 10 then a variation of strength is notified only if it is either higher or lower than 10 units of strength with respect to the last notified strength, otherwise such strength variation is

 ISBN:978-1-61208-113-7

considered irrelevant.

The graphs in Figures 1, 2 and 3 present the results obtained after applying *QuickPath* and *ShortPath* algorithms to a network with N=10 to 100 nodes.

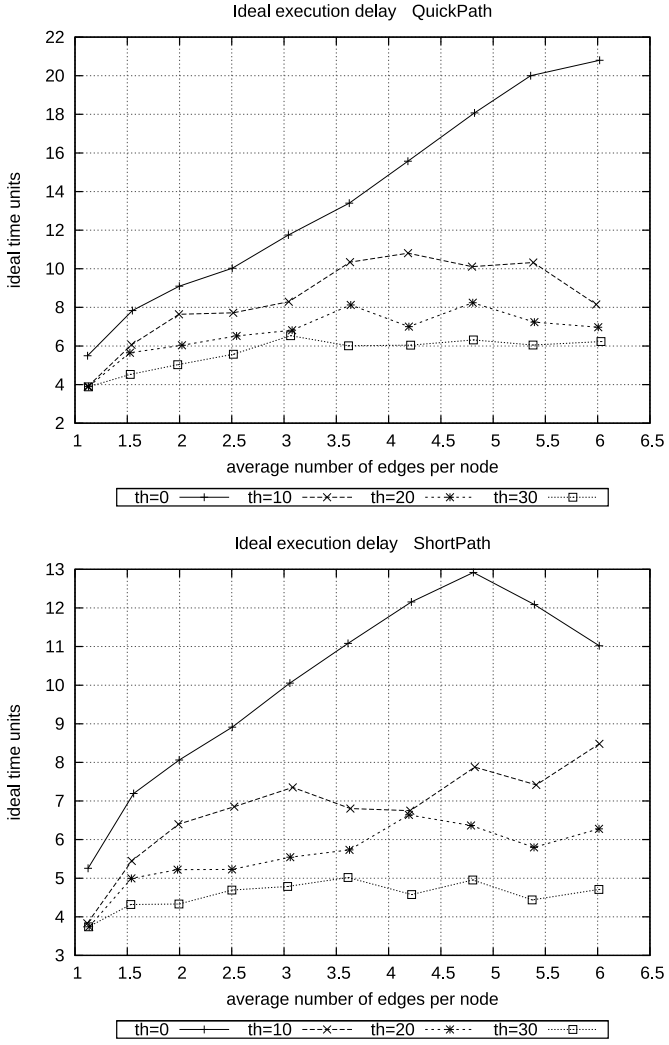


Figure 1. Ideal Execution Delay till stabilization for *QuickPath* and *ShortPath* algorithms with threshold values (th) ranging from 0 to 30.

The number of edges per node ranges from (*N-1)/N* (a tree, since the neighborhood graph is always connected) to approximately 6, as the number of nodes in the fixed-size area is increased.

Assuming that a single message takes *one time unit* to be processed and transmitted over any communication link – the "*ideal time*" [5] –, the graphs in Figure 1 permit to conclude that the total amount of ideal time units tends to stabilize as the number of edges per node grows, thus indicating excellent scalability. Moreover, this tendency is confirmed when the curve with the *threshold* value greater than zero is analyzed.

Figure 2 shows the total amount of messages exchanged by all nodes in order to update all magnetic field information in the network. Notice that the growth of the number of messages in the whole network obeys a quadratic equation as the average number of edges per node grows. However, its growth per node (not shown) can be described by a linear equation, which demonstrates again the scalability of the algorithms. Also, the number of messages issued by *QuickPath* drops faster than *ShortPath*'s as the threshold increases, since *QuickPath* decides to carry out further propagations using a single condition that is greatly influenced by the threshold value.

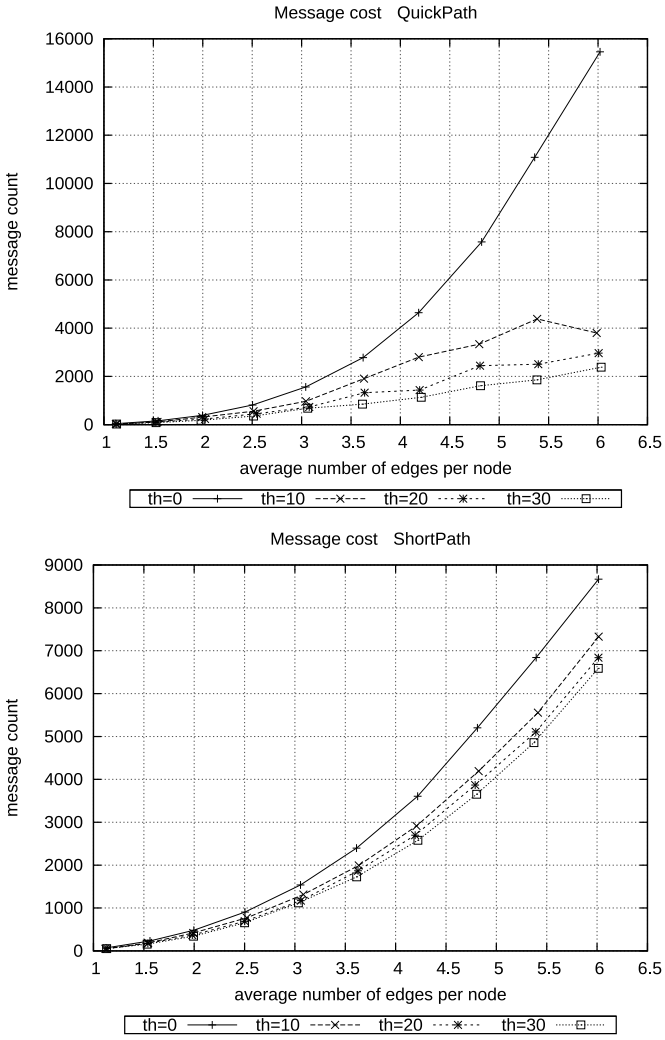


Figure 2. Comparative overall number of messages exchanged by all nodes till magnetic fields stabilization in each algorithm, considering threshold values ranging from 0 to 30.

The average number of bytes transmitted in each link is represented by the graphs in Figure 3. It can be noticed that the curves growth is close to linear, yet again indicating good scalability. Also, it should be noticed that, at instant zero, update messages are sent over *every* edge of the network while, at each time slot, up to the ideal execution delay, the number of bytes transmitted drops till it reaches zero. This happens because, at each step, only relevant information causes further propagations.

In real-world situations (those with *th*>0), the behavior of all the curves indicate good scalability. Just to have an idea of

real values, in a network with 1.5 Mbps links, the stabilization delay is approximately 6 ms in the worst case (with a maximum initial destabilization and with *th*=0).

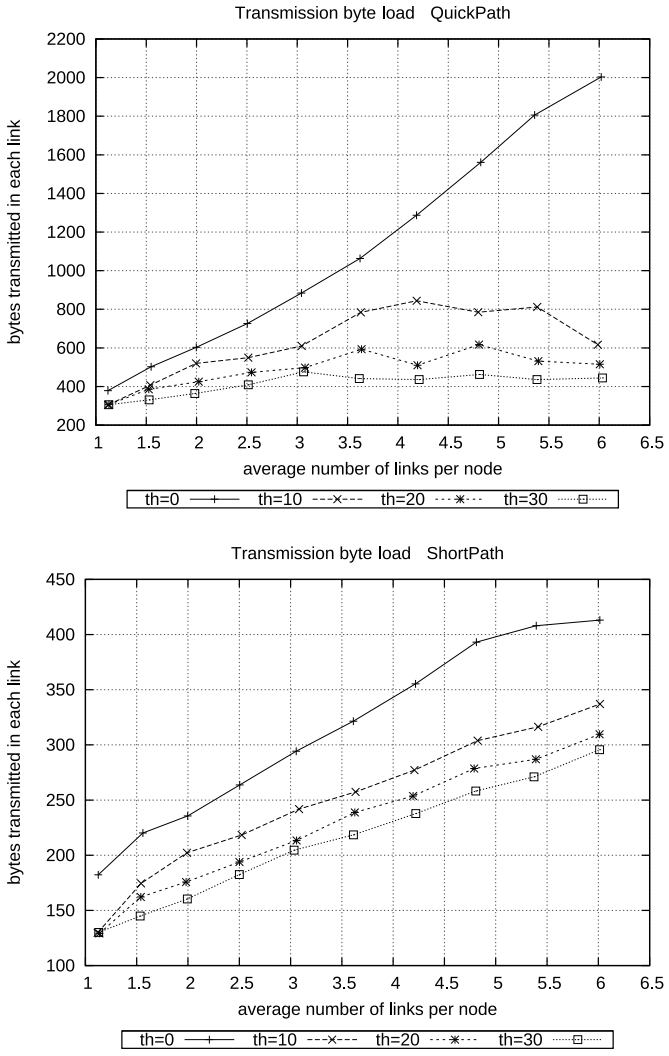


Figure 3. Average number of bytes transmitted in each link for both *QuickPath* and *ShortPath* algorithms with threshold values ranging from 0 to 30.

Other measures like local memory consumption and node workload also show similar results, i.e., linear growth, pointing to good scalability of both algorithms.

VI. CONCLUSION AND FUTURE WORK

This paper presented a novel mechanism for load balancing based on the concept of virtual magnetic fields. There is no centralized global scheduler, as tasks are simply forwarded to the idlest source node, according to the corresponding magnetic field. Also, there is no particular component to manage workload information, as every relevant change of workload in a network element is perceived by all magnetized nodes, recursively, in order to update the corresponding magnetic field. Two distinct distributed autonomic algorithms for dynamically updating magnetic fields were verified by means of simulation so that their feasibility and performance could be evaluated. The results showed that the proposed mechanism is effective, requiring acceptable cost of storage, processing and communication. In fact, the simulation results permitted to conclude that both algorithms scale very well.

The proposed mechanism can be further investigated, as follows. Firstly, the behavior of the mechanism should be verified in continuous systems operation, i.e., when tasks are started anywhere anytime, and also for the cases where nodes are heterogeneous. In such scenario, the efficiency of the proposed mechanism should be verified according to a theoretical analysis on the approximation ration to the optimum solution. Secondly, network faults should be injected to verify the correctness of the mechanism, as well as, the impact of such faults on its success rate. Thirdly, magnetic fields can be experienced in the context of cloud computing and, then, compared to standard solutions to scheduling and load balancing. The implementation of the proposed mechanism on top of a real-world platform will permit to assess its performance in terms of throughput of virtual transport connections, such as TCP and UDP. Fourthly, other important issues, such as trust, reputation and security should be considered. For example, an aspect to analyze is the effect of the transitive property of magnetization relationships on QoS and security requirements. Fifthly, the proposed mechanism can be directly compared to other load balancing techniques, such as those mentioned in Section II, and also techniques based on the swarm approach, on active networks and on mobile agents.

ACKNOWLEDGMENT

This work was partially funded by Fundação Araucária (367/2010 - Prot: 20.063) and CNPq (482593/2010-5).

REFERENCES

[1] Lo, V., Zappala, D., Zhou, D., Liu, Y., and Zhao, S. “Cluster computing on the fly: P2P scheduling of idle cycles in the Internet”. In 3rd International Workshop on Peer-to-Peer Systems (IPTPS 2004), pp. 227-236, 2004.

[2] Yagoubi, B. and Slimani, Y. “Task load balancing strategy for grid computing”, Journal of Computer Science, 3(3):186-194, 2007.

[3] Ting, Y.-W. and Chang, Y.-K. “A novel cooperative caching scheme for wireless ad hoc networks: GroupCaching”. In International Conference on Networking, Architecture and Storage (NAS 2007), pp. 62-68, 2007.

[4] Lima Jr., L.-A. and Calsavara, A. 2010. “Autonomic application-level message delivery using virtual magnetic fields”. Journal of Network and Systems Management, 18(1):97-116, March 2010.

[5] Santoro, N. “Design and Analysis of Distributed Algorithms”, Wiley, 2006, ISBN: 978-0-471-71997-7.

[6] Zhang, Q., Cheng, L., and Boutaba, R. “Cloud computing: state-of-the-art and research challenges”. Journal of Internet Services and Applications, 1(1):7-18, 2010.

[7] Randles, M., Lamb, D., and Taleb-Bendiab, A. “A comparative study into distributed load balancing algorithms for cloud computing”. In IEEE 24th International Conference on Advanced Information Networking and Applications, pp. 551-556, 2010.

[8] Sharma, S., Singh, S., and Sharma, M. “Performance analysis of load balancing algorithms”. In World Academy of Science, Engineering and Technology, 38:269-272, 2008.

[9] Calsavara, A. and Lima Jr., L.-A. “Routing based on message attraction”. In 24th Advanced Information Networking and Applications Workshops (WAINA), pp. 189-194, 2010.

 ISBN:978-1-61208-113-7

Moving to the Cloud: New Vision towards Collaborative Delivery for Open-IPTV

Emad Abd-Elrahman and Hossam Afifi
Wireless Networks and Multimedia Services Department, Telecom SudParis (ex. INT), France.
9, rue Charles Fourier, 91011 Evry Cedex, France.
{Emad.Abd_Elrahman, Hossam.Afifi}@it-sudparis.eu

Abstract*— This paper provides a short description and visions about the convergence network architecture for an Open-IPTV model based *Cloud*. Also, it counts the benefits which derived from adoption of this architecture to mobility and security issues. With the new *Open-IPTV* model, the migration towards convergence networks between different *Content Providers* (CP) infrastructures becomes evident. This adaptation leads us to the *Cloud* Computing Revolution for CPs interconnections. The traditional mobility issues will be eliminated under the new umbrella conditions for the collaboration between the CPs. Moreover, the general management methodology will add value to the overall control between those providers. It's the first step from a technical perspective, and we consider it to help in resources optimization regardless of our exact choice of any content provider. This optimization will include the two cost relations that are very important to all providers: *Capital expenditure (CAPEX)* and *Operational expenditure (OPEX).

Keywords- Collaborative Computing; Domestic Cloud; Open-IPTV

I. INTRODUCTION

Recently, the aspect of *Cloud* network has proliferated within the Internet Service Providers ISPs. It enhanced all the multimedia business efficiencies though its establishment was few years ago. *Cloud* infrastructure solutions have matured and provided reasonable interoperability under the current Internet regulations and standards. While the cost of deploying delivery networks solution for IPTV has increased over the last several years, the operational expense of maintaining and managing the network also continues to rise.

As more and more IPTV application migrate to IP Multimedia (IMS) [2] services or Next Generation Network (NGN) [1] architecture, the cost of implementations and network troubleshooting for performance issues are increasing. Unlike traditional models, the collaborative *Cloud* model poses unique challenges given the transient and shared nature of the communication medium. The ability to effectively analyze and manage problems is indispensable for maximizing the Return-on-Investment (ROI) from the *Cloud* solution. We search for significant improvement in the IPTV network delivery performance for service and content providers.

A. Open IPTV

The current status of IPTV model can be summarized in Fig. 1. We have three models; IMS based standard model based on the IP Multimedia Subsystem core as a controller and NGN based standard model based on Next Generation Network architecture and finally the more famous business Internet model which is the Google TV model [9].

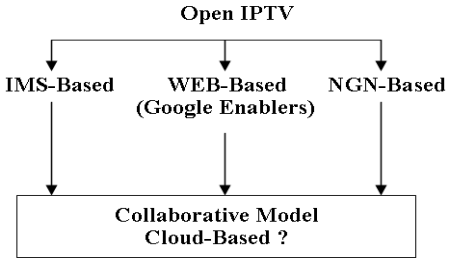


Figure 1: IPTV Models.

We expect a collaborative model (*Cloud-Based*) for IPTV delivery. This model will have some advantages over the other models in terms of low cost, good performance delivery and converged system in the domestic region.

B. Definitions

The Internet Protocol TeleVision (IPTV) term has found many definitions. But, the ITU-T definition [6] is the more intuitive one which said: "IPTV is defined as multimedia services such as television/video/audio/text/graphics/data delivered over IP based networks managed to provide the required level of QoS/QoE, security, interactivity and reliability." It is really a comprehensive definition.

In the next days, we expect a merging in IPTV technologies. Now, we mainly have three models:

- TV for normal access by traditional TVs (DVB-T).
- IPTV online channels for normal access by PCs or IP-phones.
- VoD which represents the offline case and access by same methodologies; PCs or IP-phones. The last two cases represent the web-based content.

The above three scenarios will be integrated into one scenario which represents the Future IPTV. This scenario will require a special convergence in the operator networks and some integration between different content providers.

For the General IPTV Models Management, there are two models:

- **The Managed Model**: it concerns access to and delivery of content services over an end-to-end managed network. Orange in France represents one type of this model.
- **The Unmanaged Model**: it concerns access to and delivery of content services delivered over an unmanaged network (e.g., The Internet) without any

 ISBN:978-1-61208-113-7

quality of service guarantees. YouTube represents one type of this model because there is no guarantee for QoS while accessing its videos through Internet.

This work provides a study for the current situation of IPTV delivery and how we can evolve *Cloud* starting from the current position in the access network. The main two aspects of *Cloud* computing that are interesting to our study are: *Infrastructure as a Service (IaaS)*, and *Software as a Service (SaaS)* [10]. With IaaS, a service provider delivers raw resources, like virtual machines, storage, and network bandwidth, as a service. With SaaS, a provider layers a specific software solution on top of those raw resources, and delivers it.

The rest of this work is organized as: Section II highlights some related work. In Section III, we introduce all existing terminologies related to IPTV. Section IV differentiates between the current model and our proposed one for collaborative delivery based on *Cloud* network for domestic access. Our model analysis is discussed in Section V. Section VI concludes our case study.

II. RELATED WORK

Cloud computing is a recent concept and there is few contributions in the field of collaboration in the domestic network. This is mainly due to the continuous competitive between different providers in the same region.

The concept of resources collaboration has been discussed in different places like [3]. They proposed the concept of Alliance as a general aspect of Virtual Organizations. Their Alliance concept is based on integration and collaboration between clients' requests or demands and providers resources. Moreover, they study the motivations from reforming the distinction in the current situation of organizations that will lead to good business model. The work mainly discussed the collaboration problems and some security aspects towards virtual organizations. Also, the work in [8] proposed the idea of on-demand *Cloud* service within IPTV based servers' virtualization. But, this work did not touch the area of domestic collaboration between different providers.

III. IPTV AND THE NEW OPEN BUSINESS MODEL

A. The Common IPTV Terminologies

- **Pay-TV:** this service refers to the subscription-based TV delivered in either traditional analog forms, digital or satellite. We have in different countries similar term called Packs Channels like Showtime, ART and so on.
- **TV-OTT:** TV Over-The-Top; it is one of the American modern TV term that provides a seamless consumer experience for accessing linear content through the broadcast network on a TV set, as well as non-linear services such as Catch-up TV and Video on Demand through a broadband IP network. It is also designed to allow the provider to extend content and the consumer experience to additional platforms including PCs, mobile, gaming consoles and connected TVs.
- **IPTV ''Follow-me'':** allows the user to continue access his IPTV service while moving and changing his screen. (Content Adaptation while Mobility).
- **Personal IPTV** "My Personal Content Moves with Me"): it means, allowing the user to access to his personalized IPTV content in any place in his domestic region and be billed on his own bill (*Nomadic Access*).
- **TVE:** TV Everywhere; it is the process of adding place shifting technology to (STB) set-top boxes and this required software S/W STB and also adapted (content optimization) to match all types of screens: Traditional TV, PC or laptop and Smartphone.
- **Open IPTV**: is the new model of TV service that will a borderless technology. It will be a hybrid model that merges the traditional Broadcasting TV with the Web-TV in one thing for nomadic services.

B. Physical Set-Top-Box (P-STB)

Definition:

This system represents the actual implemented scenario in the most developed countries. It mainly depends on physical *Hardware* of STB and leased connection between the consumer and content provider.

Advantages:

The service security assurance and bandwidth satisfactions are the most pros of this model. Also, the good management provided by the content providers.

Drawbacks:

With the present model of IPTV, the delivery is based on physical STB restricted to specific location. But, the consumers increasingly become *Anywhere Consumers* and they demand bandwidth regardless of their locations for satisfy their entertainments. So, the lack in this model is the inability for consumer satisfactions while changing their locations (*Mobility and Nomadic Access Aspects*). By experience, the new open model of the TV and the entire video delivery could be enhanced rapidly.

C. Software Set-Top-Box (S-STB)

Definition:

It is the new up to come system for IPTV business model. It will depend on *Software* instead of *Hardware* for controlling the received channels and videos.

Advantages:

It will satisfy the consumer desires for enjoying all their subscription videos *Anywhere*.

Drawbacks:

The operators have some fear from the management of user policies and authorizations while the clients are changing their locations.

As the future recommends the new Software model of Set-Top-Box (S-STB), we will not need to a strict physical location of IPTV services.

 ISBN:978-1-61208-113-7

D. Design Factors

Two factors press on the providers decisions while they are taking a new infrastructure investment:

- **Capital expenditure (CAPEX):** is representing the cost of network foundation and all non-consumable system devices and infrastructure.
- **Operational expenditure (OPEX):** is representing the running cost for provider network including all cost of operation and maintenance.

For the long term investments, the operators will reduce those costs in the domestic *Cloud*. Moreover, the new added services related to quality and interactivity will be costless.

IV. COLLABORATIVE ARCHITECTURE

The competitive space between different IPTV operators pushes them to implement high similarity in clients' services. This means that, the majority of VoD and IPTV channels are the same which reverse the culture and social interests of each country. Thus, if we make some convergence between the different providers it will not affect the overall policy of this country. Moreover, it will enhance the service delivery and reduce incremental cost for future service investments.

A. Current Architecture Status

The current architecture of content providers (as shown in Fig.2 as France case study) has mainly three layers:

Layer 1: Interconnection Layer (Infrastructure Providers)
Layer 2: Control Layer (Content Providers Islands)
Layer 3: Management Layer (The 3rd party hosting and cashing videos and channels servers)

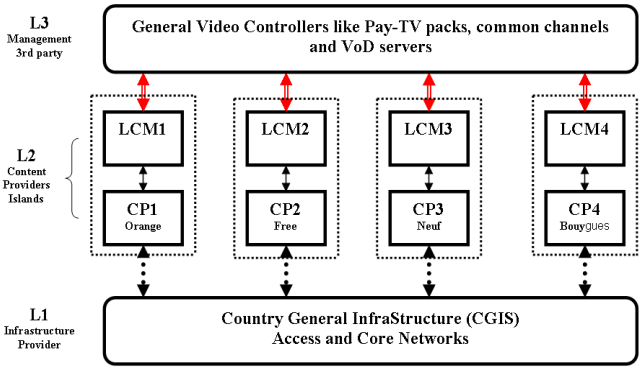


Figure 2: The current model of different islands content providers.

CP: Content Provider
LCM: Local Content Management
GCM: General Content Management
is the interface between CP access network and the core network
is the interface towards GCM and LCM

The isolations and different islands are the main features of this model. Each operator has large investments and local management for the same services provided by the others in many cases.

The open models require an open infrastructure design and also an open management policy. So, the providers must avoid their selfishness and think for two important things:

- the great benefits from the collaboration that will adopt cloud computing design
- the satisfaction of consumers towards new services

Thus, the Open-IPTV model needs a lot of cooperation between different partners for achieving remarkable success.

B. Proposed Architecture

The collaborative model design is illustrated in Fig. 3. This model proposes more interactions and collaborations between different operators. As mentioned, the S-STB is the future aspect for IPTV delivery; we use it as the point interface to multi-screen access client.

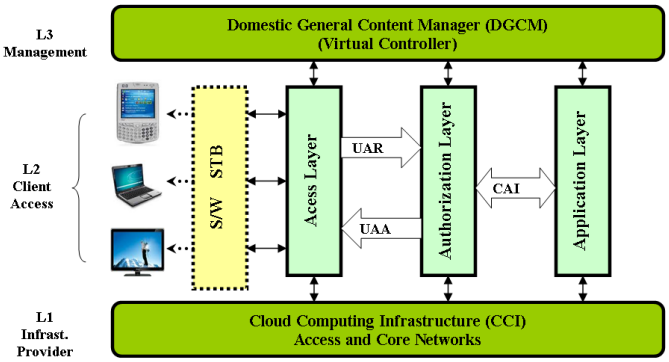


Figure 3: The proposed collaborative model for domestic cloud network to new IPTV system.

UAR: User Authorization Request
UAA: User Authorization Answer
CAI: Content Adaptions Interactions

The collaboration exists in the form of common access to CCI and DGCM layers by the client access layer. This case will lead us to new methodology of accessing which called Resource-On-Demand ROD. Moreover, this ROD will save the time and cost for service configuration. Another thing, UAR and UAA processes for clients' services authentication and authorizations pass mutually and independently of user access network. More over, the content adaptation for different screen has two aspects; one based the capability of S-STB and the other on the access device specifications. The CA process for the domestic sphere for the client is mainly done by STB.

C. Open-IPTV Model and Relation Aspects

To support a correct model, we need to explain the relations between the four billers that lead to a successful Open-IPTV model (see Fig.4) as follows:

- **Content Providers:** must study between them the content convergence so as to facilitate the consumer access methods and delivery and also the content adaptation to match different screens.
- **Operators:** must convince the delivery of Open-IPTV before missing the dominant and control of Web delivery because the service will come in the near future.
- **Infrastructure Providers:** it is the time to convergence infrastructure and *Cloud Computing* design to appear so as to enhance the user access methods and ameliorate the mobility and security user issues.

 ISBN:978-1-61208-113-7

- **Consumers:** no way for the consumers from integrating him self with the new technology and new methodologies of future Internet services. If the consumer does not conceive with this developing then we will have a missing part in the ring or the cycle will not be completed. Client culture and motivations must be changed so as to help in the model successful.

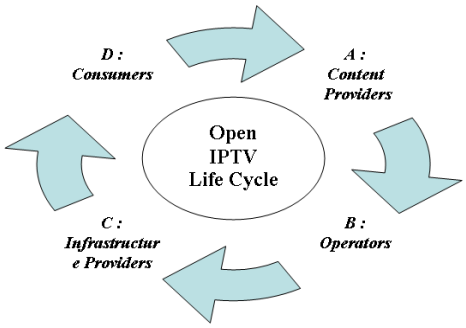


Figure 4: The Open-IPTV service billers

A: is the first point of the cycle and responsible for contents
B: is the service provider
C: is the network access part and the point of attachment for the client
D: is the last point of the cycle which represents the end user
(In some cases ***A, B, C*** can be represented by one provider)

V. THE MODEL ANALYSIS

This part provides some aspect and study analysis relevant to the new Open-IPTV model and the impacts on all members of the new life cycle as the following:

- The impacts of infrastructure integration on workflows and privacy policy.
- The impacts of applying and integrating convergence model between different content providers on:
  - consumer privacy protection measures
  - business operational cycle
  - Financial management performance.
- The optimization of resources under new collaborative conditions.
- The new behaviour of consumers under new methodologies of services accessing with different screens as shown in Fig. 5.

So, the content adaptation and the QoS assurance are the two factors which are affecting on studying multi-screen IPTV delivery and band width optimization. Moreover, these two factors are playing an important role *Cloud* migration. They are the turning point in the design and collaboration between different providers.

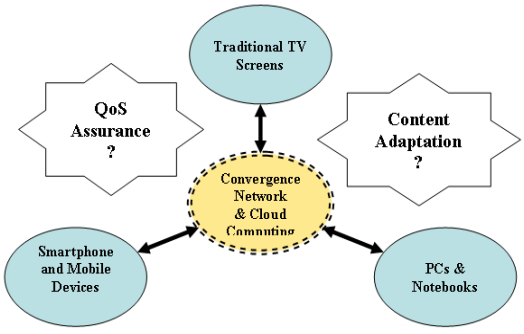


Figure 5: Open-IPTV model and Multi-screens consumer issues

A. Traditional Broadcast TV versus Web-TV

The future prospection for the relation between traditional TV watching and Web-TV in terms of average number of hours/month is shown in Fig. 6. Some IPTV weekly monitor sites [5] and specialist in technical and industrial reports [4] estimate that, the normal human in advanced countries like US and Europe watches traditional TV for 120 hours per month and Web-TV for 18 hours per month in 2010. But, this scenario will be inversed after the year 2019. They expected that, the Web-TV watching will exceed the traditional TV in 2020 and this excess will continue as shown in Fig. 6.

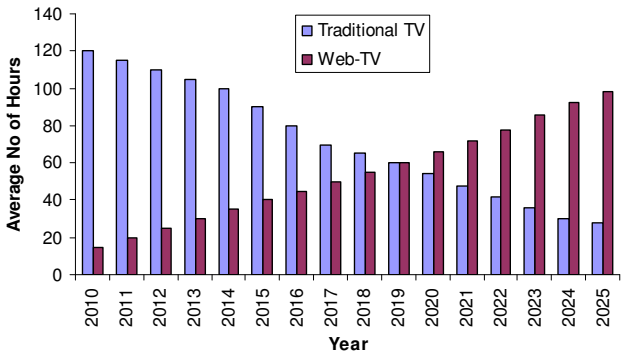


Figure 6: The expectation watching for Web-TV will exceed the amount of watching traditional TV by the year 2020.

New research shows that while TV broadcasts still dominate initial viewing, more users are turning to the new services like (TVE) to watch TV when they have time.

For some segments of the industry TVE is being able to watch any television show, any movie, any online video content anywhere you'd like on any screen, television, PC, mobile phone and, of course, the Smartphone at any time.

B. Cloud Design Motivations

Content providers businesses rely on the network (infrastructure) and the data centre (hosting servers). They are the key factor in providing a successful IPTV service. Without them, achieving business goals like increasing market share, customer satisfaction or operating margins and profitability is nearly impossible. However, over time, both networks and the data centre have matured and, in the process, become more complex than ever before. The adoption of new trends, applications and services over the years has brought with it inflexible designs, point product solutions, and a plethora of operating systems and management applications. This complexity leads to increased operating expenses in both the delivery and data centre networks and limits their potential. So, the traditional ways for enhancing the current state by hierarchal way are not suitable. They lead to more sophisticated network and high cost especially the *Capital Expenditure (CAPEX)* one. Also, by default, as a result of complexity the *Operational expenditure (OPEX)* will be augmented. So, we expect that the curve of cost will reach a high peaks.

From all causes mentioned above, we can conclude that the

 ISBN:978-1-61208-113-7

mean two reasons for adopting *Cloud Design* are the reducing of high costs of investments and consumer quality assurance for satisfying the service. But, the migration towards *Cloud* is not easy from the providers' prospection.

So, we suggest the first step of moving to the *Cloud* by adopting the convergence mechanism between the current providers. Then the future *Cloud* will be the international one. We categorize the *Clouds* into two aspects as:

- Domestic Cloud: resulted from collaborative providers in the same country.
- International Cloud: is the co-operation between different domestic clouds over Internet.

C. Collaboration Culture and Benefits

The collaboration model as a first step to cloud design is considering as a new value added culture between different providers. It will reduce the whole cost for new services deployments and increasing the total revenue from the network services. As, the future IPTV services goes to future generation based on multi-screen multi-services, the development of content and service providers must take parallel path for achieving customers satisfactions.

- High availability
- Good scalability
- Minimum OPEX cost and moderate CAPEX cost
- Good interoperability between different providers
- More facilitations to clients access services like mobility issues

On the other hand, the realization part of Fig. 4 is not practically simple. This is because the huge cost for content provider implementations to adopt multi-screen systems flow adaptations. But, we think that the collaborative model will achieve the most cost reduction for this scenario.

D. Generating Domestic Cloud

The main problem which faced all content providers toward migration to *Cloud Computing* networks is the lack in security. If the content is very sensitive, the cloud can not guarantee its security as much as required. So, we are claim to the collaboration and re-evaluation of the current infrastructure and reuse it in a *Cloud* manner through domestic area.

The reliability and troubleshooting are also representing big difference between the *Cloud* providers and traditional ones. But, the new release of *Domestic Cloud* can adopt the convergence mode.

So, the cooperation between the existing providers for obtaining a convergence network could lead them to the future *Cloud* computing network as a suitable solution in this time. The trusting in sharing contents and resources is a way toward full migration to the concept of *Cloud*. We suggest the collaborative solution as a domestic *Cloud* network for all types of videos and IPTV service.

Actually, we noticed that in France the majority of IPTV providers provide the same group of VoD, channels, Pay-TV packs and other types of videos. They almost have 90% of the contents common. Parts of these videos are hosted by another party or cashing systems like AKAMAI system [7] and small parts of videos are hosted by the providers themselves. So, if they share their resources, they will provide good services and they can overcome on the bandwidth bottleneck problems. The benefits from this convergence are:

- Interoperability: get unified management for different providers' networks.
- Portability: achieve some degree of mobility between operators.
- Integration: obtain complete service
- Quality of Services (QoS): service assurance.

The successful model of the domestic cloud will lead to host international services which are the core stone of cloud networks and *International Cloud*. The third-party cloud service is the great objective but the route we follow must be taken step-by-step.

E. CAPEX vs. OPEX Analysis

All direct and indirect infrastructures and servers costs are representing the whole part of CAPEX cost. We estimate that, for the Cloud design, it can have more reduction of the essential costs till 80% of the total cost. Moreover, the reduction mainly depends on the degree of collaboration between the providers in the same domestic region.

Let the following assumptions:

N: is the number of providers mesh links in the domestic region.

C: is the total cost for each provider.

Then, the total CAPEX for all providers = $N*C$

For collaboration Model cost ratio, it will equal: $N*c/N*C$

Where $c = L*C^2$

So,

$$CAPEXc = (L*C^2)/C = L*C$$

Where:

CAPEXc: is collaboration CAPEX cost

L: is the infrastructure Link foundation between providers after collaboration and interoperability/compatibility added costs.

So, the profit from collaboration as a reduction in CAPEX:

$$R = N*C - L*C = (N-L)*C$$

If we have a third parity (Cloud provider), then the CAPEX for the current operators will be *Zero* and all costs will just be OPEX costs.

Therefore, the typical cost optimization regarding to traditional data center design versus *Cloud* design is really remarkable. The long term costs will be reduced. Also, the benefits from adoption of *Cloud* are the utilization principles of servers based needs. This means that, the using of the infrastructure only when there is a real need and releases it for other free use times. Moreover, the Cloud is the mean of scalability for multimedia delivery.

VI. CONCLUSION

Our work contributions were to: present state-of-the-art for new IPTV terminologies like (TVE, OTT, Pay-TV and Open

TV); illustrate the benefits in convergence networks design and their impacts on CAPEX and OPEX costs. Also, demonstrate the performance of future IPTV service against traditional TV and introduce the Multi-screen idea and the bandwidth optimization for multimedia delivery to different consumer's screens and then analyze the impacts of using *Cloud* computing infrastructure in the converged network to different providers.

Finally, we expect the demise of isolated content providers' islands over the Internet at least in the domestic regions. So, the gates are now opened for clients *Nomadic Access* to *Nomadic Services (NA to NS)*.

In the next work, an investigation will be conducted for some new IPTV use cases in the domestic region using *Cloud*.

VII. ACKNOWLEDGMENT

This work is a part of feedbacks of future IPTV network delivery design in the contribution of the European project UP-TO-US (User-Centric Personalized IPTV UbiquitOus and SecUre Services) that lunched in September 2010.

REFERENCES

[1] Telecommunications and Internet converged Services and Protocols for Advanced Networking TISPAN); Service Layer Requirements to integrate NGN Services and IPTV, ETSI TS 181 016 V3.3.1, July 2009.

[2] Draft ETSI TS 182 027 V0.0.9 (2007-04), Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IPTV Architecture; IPTV functions supported by the IMS subsystem, ETSI Technical Specification Draft, 2007

[3] J.M. Brooke and M.S. Parkin; ''Enabling scientific collaboration on the Grid''; Original Research Article Future Generation Computer Systems, Volume 26, Issue 3, March 2010, pp. 521-530.

[4] ReportLinker: http://www.reportlinker.com/

[5] FierceIPTV: http://www.fierceiptv.com/

[6] ITU-T Newslog: http://www.itu.int/ITU-T/newslog/IPTV/

[7] AKAMAI: http://www.akamai.com/

[8] P. Yee Lau, S. Park, J. Yoon and J. Lee; ''Pay-As-You-Use On-Demand Cloud Service: An IPTV Case''; International Conference on Electronics and Information Engineering (ICEIE 2010); pp. V1-272 - V1-276.

[9] Google TV: http://www.google.com/tv/

[10] A.B. Bondi, 'Characteristics of scalability and their impact on performance', *Proceedings of the 2nd international workshop on Software and performance,* Ottawa, Ontario, Canada, 2000, ISBN 1-58113-195-X, pp. 195 – 203.

 ISBN:978-1-61208-113-7

Half-Band FIR Filters for Signal Compression

Pavel Zahradnik and Boris Šimák
Department of Telecommunication Engineering
Czech Technical University in Prague
Prague, Czech Republic
zahradni, simak@fel.cvut.cz

Miroslav Vlček
Department of Applied Mathematics
Czech Technical University in Prague
Prague, Czech Republic
vlcek@fd.cvut.cz

***Abstract*—An efficient design of equiripple half-band FIR filters for signal compression is presented. Solution of the approximation problem in terms of generating function and zero phase transfer function for the equiripple half-band FIR filter is shown. The equiripple half-band FIR filters are optimal in the Chebyshev sense. The closed form solution provides an efficient computation of the impulse response of the filter. One example is included.**

***Keywords*-FIR filter; half-band filter; equiripple approximation; wavelet compression;**

I. INTRODUCTION

Half-band (HB) filters are basic building blocks in wavelet analysis [1], signal compression and in multirate signal processing [2]. The only available method for designing equiripple (ER) HB finite impulse response (FIR) filters is based on the numerical McClellan - Parks program [3]. It is usually combined with a clever "Trick" [4]. Besides this, some design methods are available for almost ER HB FIR filters, e.g. [5],[6]. No general non-numerical design of an ER HB FIR filter was found in references. In our paper we are primarily concerned with the ER approximation of HB FIR filters and with the related non-numerical design procedure suitable for practical design of ER HB FIR filters. We present the generating function and the zero phase transfer function of the ER HB FIR filter. These functions give an insight into the nature of this approximation problem. Our design procedure is based on the Chebyshev polynomials of the second kind. Based on the differential equation for the Chebyshev polynomials of the second kind, we have derived formulas for an effective evaluation of the coefficients of the impulse response. We present an approximating degree equation which is useful in practical filter design. The advantage of the proposed approach over the numerical design procedures relies on the fact that the coefficients of the impulse response are evaluated by formulas. Hence the speed of the design is deterministic.

II. IMPULSE RESPONSE, TRANSFER FUNCTION AND ZERO PHASE TRANSFER FUNCTION

A HB filter is specified by the minimal passband frequency $\omega_p T$ (or maximal stopband frequency $\omega_s T$) and by the minimal attenuation in the stopband a_s [dB] (or maximal attenuation in the passband a_p [dB]). The antisymmetric behavior of its frequency response implies the relations $\omega_s T = \pi - \omega_p T$ and $10^{0.05a_p} + 10^{0.05a_s} = 1$. The goal in the filter design is to get the minimum filter length N satisfying the filter specification and to evaluate the coefficients of the impulse response of the filter. We assume the impulse response $h(k)$ with odd length $N = 2(2n+1)+1$ coefficients and with even symmetry $h(k) = h(N-1-k)$. The impulse response of the HB FIR filter with the length $N = 2(2n+1)+1$ contains $2n$ zero coefficients as follows

$$\begin{aligned} h(2n+1) &= a(0) = 0.5 \\ 2h(2n+1 \pm 2k) &= a(2k) = 0 \ , \ k = 1 \dots n \\ 2h(2n+1 \pm (2k+1)) &= a(2k+1) \ , \ k = 0 \dots n \ . \end{aligned} \tag{1}$$

The transfer function of the HB FIR filter is

$$H(z) = z^{-(2n+1)} \left[\frac{1}{2} + \sum_{k=0}^{n} a(2k+1)\, T_{2k+1}(w) \right] \tag{2}$$

where $T_m(w)$ is Chebyshev polynomials of the first kind. The frequency response $H(e^{j\omega T})$ of the HB FIR filter is

$$H(e^{j\omega T}) = e^{-j(2n+1)\omega T}\ Q(\cos \omega T) \tag{3}$$

where $Q(w)$ is a polynomial in the variable $w = (z + z^{-1})/2$ which on the unit circle reduces to a real valued zero phase transfer function $Q(w)$ of the real argument $w = \cos(\omega T)$.

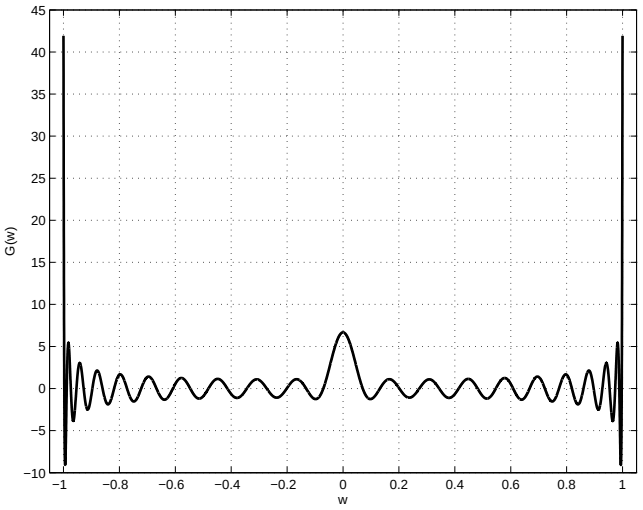


Fig. 1. Generating polynomial $G(w)$ for $n = 20$, $\kappa' = 0.03922835$, $A = 1.08532371$ and $B = 0.95360863$.

 ISBN:978-1-61208-113-7

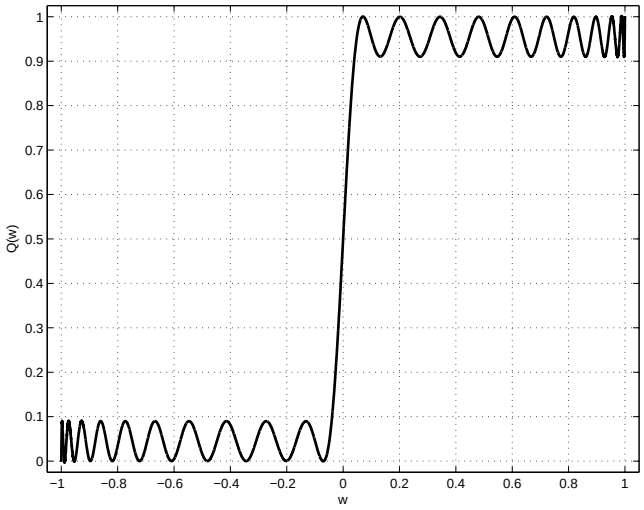


Fig. 2. Zero phase transfer function $Q(w)$ for $n = 20$, $\kappa' = 0.03922835$, $A = 1.08532371$, $B = 0.95360863$ and $\mathcal{N}$= 0.55091994.

III. Generating Polynomial and Zero Phase Transfer Function of an ER HB FIR Filter

A straightforward theory for the generating polynomial of an ER HB FIR filter is currently unavailable. The generating polynomial of an ER HB FIR filter is related to the generating polynomial of the almost ER HB FIR filter presented in [5]. Based on our experiments conducted in [5], we have found that the generating polynomial $G(w)$ (Fig. 1) of the ER HB FIR filter is obtained by weighting of Chebyshev polynomials in the generating polynomial of the almost ER HB FIR filter, namely

$$G(w) = AU_n\left(\frac{2w^2 - 1 - \kappa'^2}{1 - \kappa'^2}\right) + BU_{n-1}\left(\frac{2w^2 - 1 - \kappa'^2}{1 - \kappa'^2}\right) \quad (4)$$

where $U_n(x)$ and $U_{n-1}(x)$ are Chebyshev polynomials of the second kind and A, B, κ' are real numbers. The zero phase transfer function $Q(w)$ (Fig. 2) is related to the generating polynomial

$$Q(w) = \frac{1}{2} + \frac{1}{\mathcal{N}}\int G(w)dw \quad (5)$$

where the norming factor $\mathcal{N}$ is given by (17). The generating polynomial $G(w)$ and the zero phase transfer function $Q(w)$ show the nature of the approximation of an ER HB FIR filter.

IV. Differential Equation and Impulse Response of an ER HB FIR Filter

The Chebyshev polynomial of the second kind $U_x(w)$ fulfils the differential equation

$$(1 - x^2)\frac{d^2U_n(x)}{dx^2} - 3x\frac{dU_n(x)}{dx} + n(n+2)U_n(x) = 0 \ . \quad (6)$$

Using substitution

$$x = \left(\frac{2w^2 - 1 - \kappa'^2}{1 - \kappa'^2}\right) \quad (7)$$

we get the differential equation (6) in the form

$$w(w^2 - \kappa'^2)\left[(1 - w^2)\frac{d^2U_n(w)}{dw^2} - 3w\frac{dU_n(w)}{dw}\right] + \left[\kappa'^2(1 - w^2) + 2w^2(1 - w^2)\right]\frac{dU_n(w)}{dw} + 4w^3n(n+2)U_n(w) = 0 \ . \quad (8)$$

Based on the differential equation (8), we have derived the non-numerical procedure for the evaluation of the impulse response $h_n(k)$ corresponding to polynomial $\mathcal{U}_n(w)$

$$\mathcal{U}_n(w) = \int U_n\left(\frac{2w^2 - 1 - \kappa'^2}{1 - \kappa'^2}\right)dw \ . \quad (9)$$

This procedure is summarized in Tab. I. The impulse response $h(k)$ of the ER HB FIR filter is

$$h(k) = \frac{1}{2} + \frac{A}{\mathcal{N}}h_n(k) + \frac{B}{\mathcal{N}}h_{n-1}(k) \ . \quad (10)$$

The non-numerical evaluation of the impulse response $h(k)$ is essential in the practical filter design.

V. Degree of an ER HB FIR Filter

The exact degree formula is not available. In the practical filter design, the degree n can be obtained with excellent accuracy from the specified minimal passband frequency $\omega_p T$ and from the minimal attenuation in the stopband a_s [dB] using the approximating degree formula

$$n \doteq \frac{a_s[dB] - 18.18840664\,\omega_p T + 33.64775300}{18.54155181\,\omega_p T - 29.13196871} \ . \quad (11)$$

The exact relation between the minimal attenuation in the stopband a_s [dB], the minimal passband frequency $\omega_p T$ and the degree n were obtained experimentally.

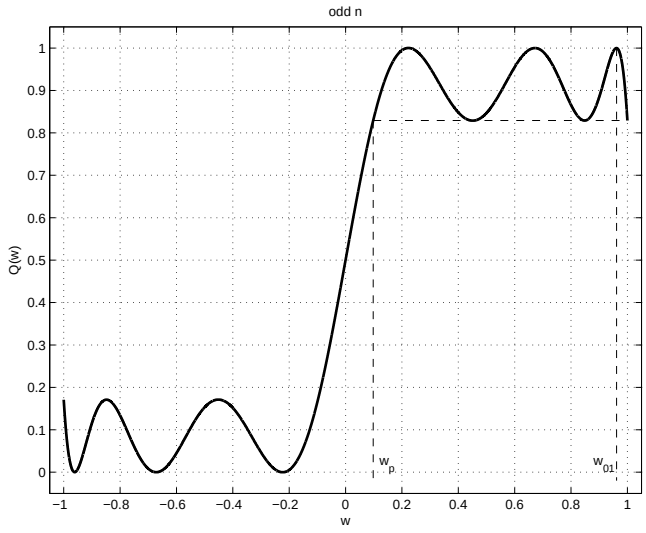


Fig. 3. $Q(w)$ for odd n.

 ISBN:978-1-61208-113-7

TABLE I
ALGORITHM FOR THE EVALUATION OF THE COEFFICIENTS $h_n(k)$.

given	n (integer value), κ' (real value)
initialization	$\alpha(2n) = \dfrac{1}{(1-\kappa'^2)^n}$
	$\alpha(2n-2) = -(2n\kappa'^2+1)\,\alpha(2n)$
	$\alpha(2n-4) = -\dfrac{4n+1+(n-1)(2n-1)\kappa'^2}{2n}\,\alpha(2n-2) - \dfrac{(2n+1)(n+1)\kappa'^2}{2n}\,\alpha(2n)$
body	
(*for* $k = n$ *down to* 3)	$\alpha(2k-6) =$
	$\{ -\left[3(n(n+2)-k(k-2))+2k-3+2(k-2)(2k-3)\kappa'^2\right]\alpha(2k-4)$
	$-\left[3(n(n+2)-(k-1)(k+1))+2(2k-1)+2k(2k-1)\kappa'^2\right]\alpha(2k-2)$
	$-[n(n+2)-(k-1)(k+1)]\,\alpha(2k)\quad\}\ /\ [n(n+2)-(k-3)(k-1)]$
(*end loop on* k)	
integration	
(*for* $k = 0$ *to* n)	$a(2k+1) = \dfrac{\alpha(2k)}{2k+1}$ (*end loop on* k)
impulse response $h_n(k)$	
	$h_n(2n+1) = 0$
(*for* $k = 0$ *to* n)	$h_n(2n+1\pm(2k+1)) = \dfrac{a(2k+1)}{2}$ (*end loop on* k)

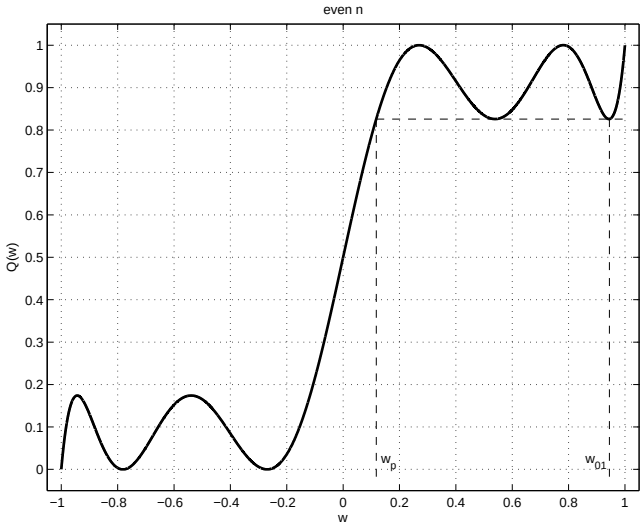


Fig. 4. $Q(w)$ for even n.

VI. SECONDARY VALUES OF THE ER HB FIR FILTER

The secondary real values κ', A and B can be obtained from the specified passband frequency $\omega_p T$ and from the degree n of the generating polynomial. In practical filter design, the approximating formulas

$$\kappa' = \frac{n\omega_p T - 1.57111377\,n + 0.00665857}{-1.01927560\,n + 0.37221484} \tag{12}$$

$$A = \left(0.01525753\,n + 0.03682344 + \frac{9.24760314}{n}\right)\kappa' + 1.01701407 + \frac{0.73512298}{n} \tag{13}$$

and

$$B = \left(0.00233667\,n - 1.35418408 + \frac{5.75145813}{n}\right)\kappa' + 1.02999650 - \frac{0.72759508}{n} \tag{14}$$

obtained experimentally provide excellent accuracy. Further, the exact values κ', A and B can be obtained numerically (e.g. using the Matlab function $fminsearch$) by satisfying the equality (see Fig. 3 - 4)

$$Q(w_p) = \begin{cases} Q(1) & \text{if } n \text{ is odd} \\ Q(w_{01}) & \text{if } n \text{ is even} \end{cases}. \tag{15}$$

The value

$$w_{01} = \sqrt{\kappa'^2 + (1-\kappa'^2)\cos^2\frac{\pi}{2n+1}} \tag{16}$$

was introduced in [6]. Relation (15) guarantees the equiripple behaviour of the generating polynomial $Q(w)$.

VII. DESIGN OF THE ER HB FIR FILTER

The design procedure is as follows:

- Specify the ER HB FIR filter by the minimal passband frequency $\omega_p T$ and by the minimal attenuation in the stopband a_s [dB].
- Calculate the integer degree n of the generating polynomial (11).
- Calculate the real values κ' (12), A (13) and B (14).
- Evaluate the partial impulse responses $h_n(k)$ and $h_{n-1}(k)$ (Tab. I).

 ISBN:978-1-61208-113-7

- Evaluate the final impulse response $h(k)$ (10) where the real norming factor $\mathcal{N}$ is

$$\mathcal{N} = \begin{cases} 2\,[\,A\mathcal{U}_n(1) + \mathrm{B}\mathcal{U}_{n-1}(1)\,] & \text{if n is even} \\ 2\,[\,A\mathcal{U}_n(w_{01}) + \mathrm{B}\mathcal{U}_{n-1}(w_{01})\,] & \text{if n is odd .} \end{cases} \quad (17)$$

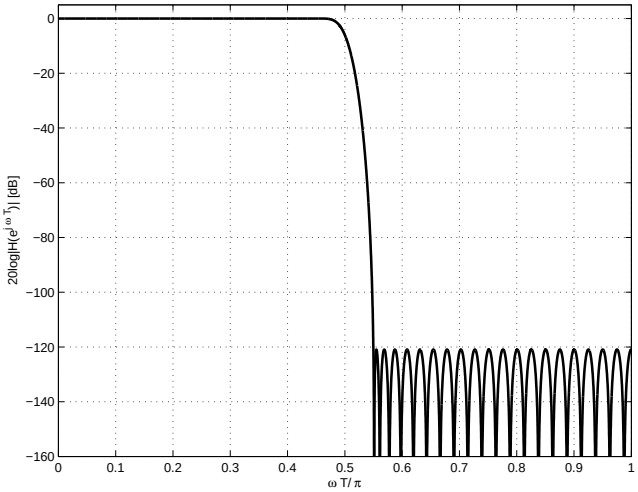


Fig. 5. Amplitude frequency response $20\log|H(e^{j\omega T})|$ [dB].

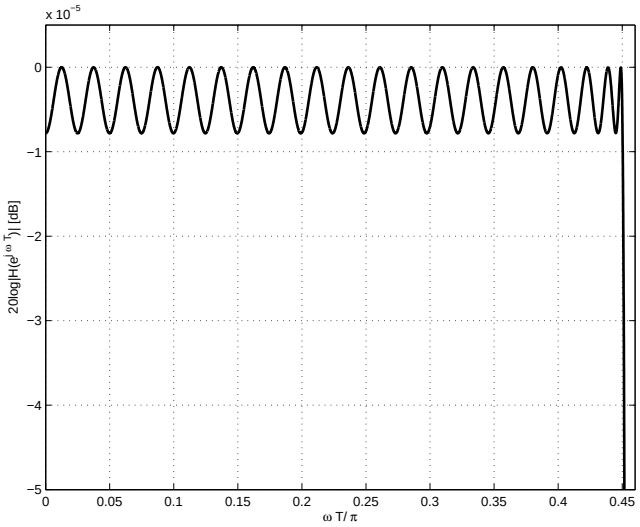


Fig. 6. Passband of the filter.

VIII. Example

Design the ER HB FIR filter specified by the minimal passband frequency $\omega_p T = 0.45\pi$ and by the minimal attenuation in the stopband $a_s = -120$ dB.

We get $n = 38.3856 \rightarrow 39$ (11), $\kappa' = 0.15571103$ (12), $A = 1.17117396$ (13), $B = 0.83763199$ (14) and $\mathcal{N} = -2747.96038544$ (17). The impulse response $h(k)$ (Tab. II) with the length $N = 159$ coefficients is evaluated using Tab. I and eq. (10). The actual values $\omega_{p\ act}T = 0.4502\pi$ and $a_{act} = -120.91$ dB satisfy the filter specification. The amplitude frequency response $20\log|H(e^{j\omega T})|$ [dB] of the filter is shown in Fig. 5. The detailed view of its passband is shown in Fig. 6. For the specified values $\omega_p T = 0.45\pi$ and $N = 159$, the comparative numerical design based on the "Trick" [3] combined with the Remez algorithm using the Matlab function $firpm$ results in the slightly unsatisfactory minimal passband frequency $\omega_{p\ act}T = 0.44922001\pi < 0.45\pi$ and consequently in a slightly better minimal attenuation in the stopband $a_{s\ act} = -123.29066608$ [dB].

TABLE II
Coefficients of the Impulse Response.

k	$h(k)$	k	$h(k)$
0 , 158	-0.00000070	42 , 116	0.00231877
2 , 156	0.00000158	44 , 114	-0.00283354
4 , 154	-0.00000331	46 , 112	0.00344038
6 , 152	0.00000622	48 , 110	-0.00415347
8 , 150	-0.00001087	50 , 108	0.00498985
10 , 148	0.00001799	52 , 106	-0.00597048
12 , 146	-0.00002852	54 , 104	0.00712193
14 , 144	0.00004363	56 , 102	-0.00847897
16 , 142	-0.00006481	58 , 100	0.01008867
18 , 140	0.00009384	60 , 98	-0.01201717
20 , 138	-0.00013287	62 , 96	0.01436125
22 , 136	0.00018446	64 , 94	-0.01726924
24 , 134	-0.00025161	66 , 92	0.02098117
26 , 132	0.00033779	68 , 90	-0.02591284
28 , 130	-0.00044697	70 , 88	0.03285186
30 , 128	0.00058370	72 , 86	-0.04348979
32 , 126	-0.00075311	74 , 84	0.06223123
34 , 124	0.00096097	76 , 82	-0.10523903
36 , 122	-0.00121375	78 , 80	0.31802058
38 , 120	0.00151871	79	0.50000000
40 , 118	-0.00188398		

IX. Conclusion

This paper has presented the equiripple approximation of halfband FIR filters. The generating polynomial and the zero phase transfer function illustrate the nature of the approximation problem. The coefficients of the impulse response are straightforwardly evaluated from the filter specification.

Acknowledgement

This activity was supported by the grant No. MSM6840770014, Ministry of Education, Czech Republic.

References

[1] D. F. Walnut, *An Introduction to Wavelet Analysis*, Birkhäuser-Boston, 2002.

[2] N. J. Fliege, Multirate Digital Signal Processing, John Wiley and Sons, New York, 1994.

[3] J.H. McClellan, T. W. Parks and L. R. Rabiner, A Computer Program for Designing Optimum FIR Linear Phase Digital Filters, *IEEE Trans. Audio Electroacoust.*, Vol. AU - 21, Dec. 1973, pp. 506 - 526.

[4] P. P. Vaidyanathan and T. Q. Nguyen, A "TRICK" for the Design of FIR Half-Band Filters, *IEEE Transactions on Circuits and Systems*, Vol. CAS - 34, No. 3, March 1987, pp. 297 - 300.

[5] P. N. Wilson and H. J. Orchard, A Design Method for Half-Band FIR Filters, *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications*, Vol. CAS - 46, No. 1, January 1999, pp. 95 - 101.

[6] P. Zahradnik, M. Vlček and R. Unbehauen, Almost Equiripple FIR Half-Band Filters, *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications*, Vol. CAS - 46, No. 6, June 1999, pp. 744 - 748.

 ISBN:978-1-61208-113-7

Comb Filters for Communication Technology

Pavel Zahradnik and Boris Šimák
Department of Telecommunication Engineering
Czech Technical University in Prague
Prague, Czech Republic
zahradni, simak@fel.cvut.cz

Miroslav Vlček
Department of Applied Mathematics
Czech Technical University in Prague
Prague, Czech Republic
vlcek@fd.cvut.cz

***Abstract*—An extension of the design of digital equiripple comb FIR filters is presented. We introduce the fifth type of comb FIR filter which complements the existing four standard types. The design runs from the filter specifications through the degree formula to the impulse response coefficients, which are evaluated by an efficient recursive algorithm. One example is included.**

***Keywords*-comb filter; FIR filter; recursive algorithm; fast design; robustness;**

I. INTRODUCTION

Comb FIR filters are often used in the digital processing of communication signals. They are mainly used in the suppression of unwanted spectral components in the signal, e.g. for the attenuation of induced power line interferences. The standard design of a comb FIR filter starts with a prototype filter which is usually a low pass or high pass filter. The comb FIR filter is obtained by inserting of a finite number of zeros between the values of the impulse response of the prototype [1]. In [2], [3] we have demonstrated an efficient design of optimal equiripple comb FIR filters which outperforms the standard design both in terms of speed and robustness. All the design procedures [1] - [3] lead to the comb FIR filters of four types (Fig. 1) with available notch frequencies summarized in Tab. I. In this paper we introduce the design of comb FIR filter of another type (Fig. 2). We call it comb5 FIR filter. The comb5 FIR filter cannot be designed using design procedures [1] - [3]. The reason for this is that the frequency response of the comb5 FIR filter is not obtained by the repetition of a LP or HP prototype filter resulting from the interleaving of the impulse response of the prototype filter by zeros. Moreover, it cannot be derived from the types 1-4, e.g. using frequency transformation. The organization of the paper is as follows. In Section II the impulse response of the filter is defined. In Section III we summarize available positions of notch frequencies. In Section IV the idea of the comb5 FIR filter is introduced. In Section V we present the design of the comb5 FIR filter. Finally, Section VI illustrates the practical design of the comb5 FIR filter.

II. IMPULSE RESPONSE

We assume the impulse response $h(k)$ with odd length $N = 2\nu + 1$ coefficients and with even symmetry

$$a(0) = h(\nu) \ , \ a(k) = 2h(\nu + k) = 2h(\nu - k) \, , \, k = 1 \ldots \nu \, . \tag{1}$$

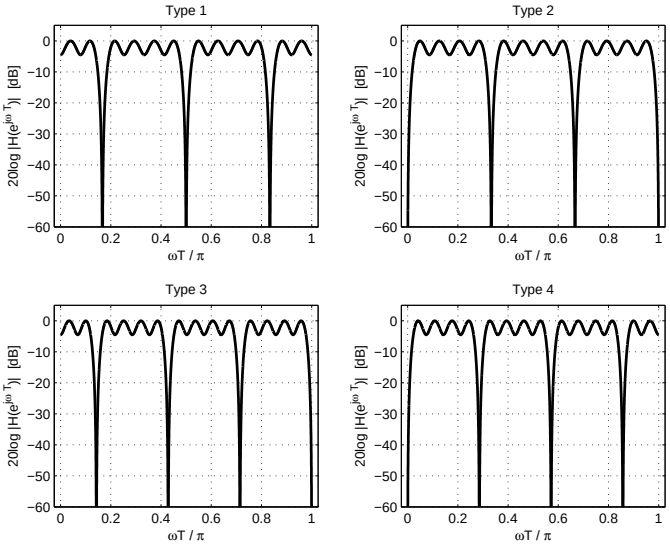


Fig. 1. Four types of comb FIR filter [3].

TABLE I
NOTCH FREQUENCIES.

Type 1 r even Marginal : none Non-marginal : $\omega_i = (2i+1)\frac{\pi}{r}$ $i = 0, \ldots, \frac{r}{2} - 1$	Type 2 r even Marginal : $\omega = 0, \omega = \pi$ Non-marginal : $\omega_i = 2i\frac{\pi}{r}$ $i = 1, \ldots, \frac{r}{2} - 1$
Type 3 r odd Marginal : $\omega = \pi$ Non-marginal : $\omega_i = (2i+1)\frac{\pi}{r}$ $i = 0, \ldots, \frac{r-3}{2}$	Type 4 r odd Marginal : $\omega = 0$ Non-marginal : $\omega_i = 2i\frac{\pi}{r}$ $i = 1, \ldots, \frac{r-1}{2}$

The transfer function of the filter is

$$H(z) = \sum_{k=0}^{2\nu} h(k)\, z^{-k} = z^{-\nu}\left[h(\nu) + 2\sum_{k=1}^{\nu} h(\nu \pm k)\frac{\left(z^{k} + z^{-k}\right)}{2}\right]$$
$$= z^{-\nu}\sum_{k=0}^{\nu} a(k)\, T_k(w) = z^{-\nu} Q(w) \tag{2}$$

where $T_n(x)$ is Chebyshev polynomial of the first kind

$$T_n(x) = \begin{cases} \cos\left(n \arccos(x)\right) & \text{if} |x| \leq 1 \\ \cosh\left(n \operatorname{arccosh}(x)\right) & \text{otherwise} \end{cases} . \tag{3}$$

 ISBN:978-1-61208-113-7

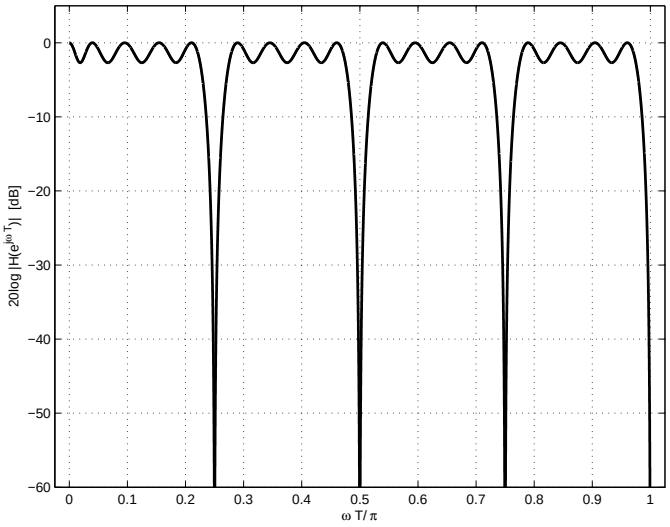


Fig. 2. Comb5 FIR filter.

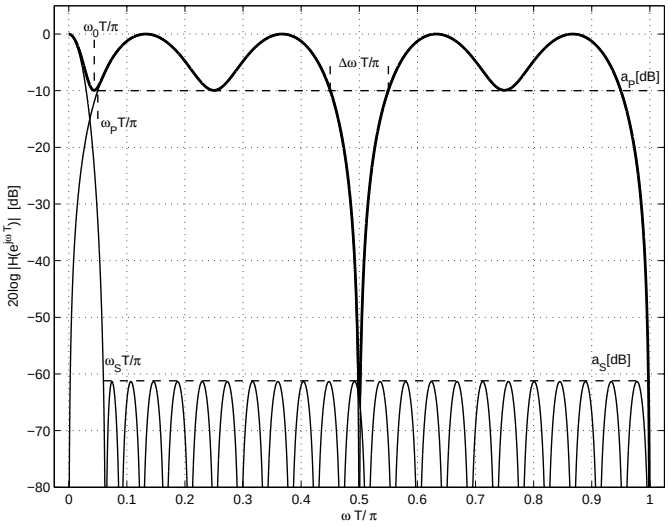


Fig. 3. Forming of the frequency response of comb5 FIR filter.

The function $Q(w)$ represents a polynomial in the variable $w = \frac{1}{2}(z + z^{-1})$ which on the unit circle $z = e^{j\omega T}$ reduces to the real valued zero phase transfer function $Q(w)$ of the real argument

$$w = \cos(\omega T)\,. \tag{4}$$

III. Notch Frequencies

There are two types of notch bands. First, the marginal notch bands are located at the notch frequencies $\omega T = 0$ and $\omega T = \pi$. The width of the marginal notch band(s) is $\Delta\omega T/2$. Second, the non-marginal notch bands are located inside the frequency interval $(0, \pi)$. The width of the non-marginal notch band(s) is $\Delta\omega T$. Four types of comb FIR filters are generated containing either none, or one, or two marginal notch bands (Fig. 1, Tab. I). The comb5 FIR filter (Fig. 2) introduced here consists of one marginal notch band at $\omega T = \pi$ and of non-marginal notch band(s) inside the interval $(0, \pi)$.

IV. The Comb5 FIR Filter

The comb5 FIR filter (Fig. 2) is in fact a comb FIR filter of type 2 with compensated attenuation at and near the frequency $\omega T = 0$. Its frequency behavior is obtained by the parallel combination of a comb FIR filter of the type 2 and of a DC-pass FIR filter which is an inverted DC-notch FIR filter [4]. In the discrete time domain, the impulse response $h(k)$ of the comb5 FIR filter is the sum of the impulse response $h_{C2}(k)$ of the comb FIR filter of the type 2 and of the impulse response $h_{DC}(k)$ of the DC-pass FIR filter. Consequently the design of the comb5 FIR filter is a co-design of the two partial FIR filters. The approximation theory behind the partial filters was treated in [3] and [4] in detail. Here we are focused on the practical aspects of the design.

V. Design Procedure

The design procedure of the comb5 FIR filter consists of two basic steps. In the first step we design the comb FIR filter of the type 2 with the zero phase transfer function [3]

$$Q(w) = 1 - \frac{1 + (-1)^n T_n\left[\dfrac{T_r(w) - \kappa^2}{1 - \kappa^2}\right]}{1 + (-1)^n \cosh\left(n\,\text{acosh}\dfrac{\kappa^2 + 1}{\kappa^2 - 1}\right)}\,. \tag{5}$$

The impulse response $h_{C2}(k)$ corresponding to the zero phase transfer function (5) consists of $N = 2rn + 1$ coefficients. The design of the comb FIR filter of the type 2 proceeds as follows:

- Specify the number b of non-marginal notch bands, the width of the non-marginal notchbands $\Delta\omega T$, the maximal attenuation in the pass bands a_p [dB] and the minimal attenuation at notch frequencies a_s [dB] (Fig. 3).
- Evaluate the degree n of the outer Chebyshev polynomial $T_n()$ in (5)

$$n \geq \frac{\text{acosh}\dfrac{1 + 10^{0.05a_p[dB]}}{1 - 10^{0.05a_p[dB]}}}{\text{acosh}\dfrac{1 + \kappa^2}{1 - \kappa^2}} \tag{6}$$

  where κ^2 is

$$\kappa^2 = \frac{1 - \cos\left(r\dfrac{\Delta\omega T}{2}\right)}{1 + \cos\left(r\dfrac{\Delta\omega T}{2}\right)} \tag{7}$$

  and

$$r = 2(b + 1)\ . \tag{8}$$

- Evaluate the impulse response $h_{c1}(k)$ of the comb filter of the type 1 (Tab. II).
- Convert the impulse response $h_{c1}(k)$ of the comb FIR filter of type 1 into the impulse response $h_{c2}(k)$ of the comb FIR filter of type 2 using matrix multiplication

 ISBN:978-1-61208-113-7

TABLE II
EVALUATION OF THE IMPULSE RESPONSE $h_{C1}(k)$.

given n, r (integer values) , $\kappa^2 \neq 0$ (real value)

initialization $a(n) = \dfrac{1}{(1-\kappa^2)^n}$, $a(n+k) = 0$ for $k = 1, 2, 3$, $h(k) = 0$ for $k = 0, \ldots, 2nr$

body

(*for* $k = 1$ ... n)

$$a(n-k) =$$

$$\begin{aligned}\{ \quad & -\frac{1}{2}\left[(2n-k+1)(k-1) + \kappa^2(n+1-k)(2(n-k)+1)\right]\ a(n+1-k) \\ & -\kappa^2(n+2-k)\ a(n+2-k) \\ & +\frac{1}{2}\left[(2n-k+3)(k-3) + \kappa^2(n+3-k)(2(n-k)+7)\right]\ a(n+3-k) \\ & +\frac{1}{4}(2n+4-k)(k-4)\ a(n+4-k) \quad \Big\}\ /\ \frac{1}{4}k(2n-k)\end{aligned}$$

(*end loop on* k)

$$a(0) = \frac{a(0)}{2} , \quad C = \cosh\left(n \cosh \frac{\kappa^2+1}{\kappa^2-1}\right)$$

impulse response

$$h_{C1}(rn) = 1 - \frac{1 + (-1)^n a(0)}{1 + (-1)^n C}$$

(*for* $k = 1$ ... n) $\quad h_{C1}\left(r(n \pm k)\right) = -(-1)^n \dfrac{1}{2} \dfrac{a(k)}{1 + (-1)^n C}$ $\quad$ (*end loop on* k)

TABLE III
EVALUATION OF THE IMPULSE RESPONSE $h_{DC}(k)$.

given $n = n_{DC}$ (integer value), λ (real value)

initialization $A(n) = \lambda^n$, $A(n+1) = A(n+2) = A(n+3) = 0$

body

(*for* $k = 2$ ... $n+1$)

$$A(n+1-k) =$$

$$\begin{aligned}\{ \quad & 2\left[(k-1)(2n+1-k) - (\lambda'/\lambda)(n+1-k)(2n+1-2k)\right] A(n+2-k) \\ & +4\,(\lambda'/\lambda)(n+2-k)\ A(n+3-k) \\ & -2\left[(k-3)(2n+3-k) - (\lambda'/\lambda)(n+3-k)(2n+7-2k)\right]\ A(n+4-k) \\ & +(k-4)(2n+4-k)\ A(n+5-k) \quad \}\ /\ k(2n-k)\end{aligned}$$

(*end loop on* k)

impulse response

$$h_{DC}(n_{DC}) = \frac{\dfrac{A(0)}{2} + 1}{T_n(2\lambda - 1) + 1}$$

(*for* $k = 1$ ... n) $\quad h_{DC}(n_{DC} \pm k) = \dfrac{1}{2} \dfrac{A(k) + 1}{T_n(2\lambda - 1) + 1}$ $\quad$ (*end loop on* k)

 ISBN:978-1-61208-113-7

$$h_{c2}(k) = \sum_{m=0}^{2rn} \Lambda(k-rn, l-rn)\, h_{c1}(m)\,,\ k = 0, \ldots, 2rn \tag{9}$$

using the matrix where

$$\Lambda(x,y) = \begin{cases} \cos\left(y\frac{\pi}{r}\right) & \text{if } x = y \\ \dfrac{2\sin\left(y\frac{\pi}{r}\right)}{\pi(y-x)} & \text{if } x \neq y \text{ and } |x-y| \text{ is even} \\ 0 & \text{otherwise} \quad . \end{cases} \tag{10}$$

In the second step, we design the DC-pass FIR filter based on the zero phase transfer function [4]

$$Q_{DC}(w) = 1 - \frac{T_{n_{DC}}(\lambda\, w + \lambda - 1) + 1}{T_{n_{DC}}(2\lambda - 1) + 1} \quad . \tag{11}$$

An important issue is the proper choice of the stop-band frequency $\omega_s T$ of the DC-pass FIR filter. This choice guarantees the value of the ripple at $\omega_0 T$ equal to the values of remaining ripples in the pass bands of the comb5 FIR filter (Fig. 3). We have no formula for the calculation of the value $\omega_s T$. Consequently we calculate the value $\omega_s T$ numerically by minimizing of the function

$$\left| |H_{C2}(e^{j\omega_0 T})| + |H_{DC}(e^{j\omega_0 T}) - a_p| \right| \quad . \tag{12}$$

The design of the DC-pass FIR filter continues as follows:

- Calculate the real parameter λ

$$\lambda = \frac{1}{1 - \sin^2\frac{\omega_s T}{2}} \quad . \tag{13}$$

- Evaluate the degree n_{DC} of the zero phase transfer function (11)

$$n_{DC} \geq \frac{\operatorname{acosh}\frac{1 + 10^{0.05 a_s[dB]}}{1 - 10^{0.05 a_s[dB]}}}{\operatorname{acosh}\frac{1 + \sin^2\frac{\omega_s T}{2}}{1 - \sin^2\frac{\omega_s T}{2}}} \quad . \tag{14}$$

- Evaluate the impulse response $h_{DC}(k)$ of the length $2n_{DC}+1$ coefficients of the DC-pass FIR filter (Tab. III).
- Finally, calculate the impulse response $h(k)$ of the comb5 FIR filter

$$h(k) = h_{C2}(k) + h_{DC}(k) \quad . \tag{15}$$

VI. Example of the Design

Design a comb5 FIR filter with 9 non-marginal notch bands of the width $\Delta\omega T = \pi/100$, maximal attenuation in the pass bands $a_p = -3$ dB and minimal attenuation in the stop bands $a_s = -60$ dB.

We get $r = 20$ using (Tab. I) and (8), $\kappa^2 = 0.02508563$ (7), $n = 7.67506483 \rightarrow n = 8$ (6). Using the recursive algorithm (Tab. II) we get the impulse response $h_{C1}(k)$ of the equiripple comb FIR filter of type 1. The impulse response $h_{C2}(k)$ of the equiripple comb FIR filter of type 2 is obtained from $h_{C1}(k)$ using (9), (10). The length of the impulse response $h_{C2}(k)$ is 321 coefficients. The actual attenuation in the pass bands is $a_{p\ act} = -2.7029$ dB. The amplitude frequency response $20\log|H(e^{j\omega T})|$ [dB] of the comb FIR filter of type 2 is shown in Fig. 4. The value $\omega_s T = 0.00668686\pi$ of the DC-pass FIR filter was evaluated numerically (12). Further, we get $\lambda = 1.00067186$ (13) and $n_{DC} = 394.8084 \rightarrow n = 395$ (14). Using the recursive algorithm (Tab. III) we get the impulse response $h_{DC}(k)$ of the DC-pass FIR filter. Its length is 791 coefficients. The amplitude frequency response $20\log|H(e^{j\omega T})|$ [dB] of the DC-pass FIR filter is shown in Fig. 5. The final impulse

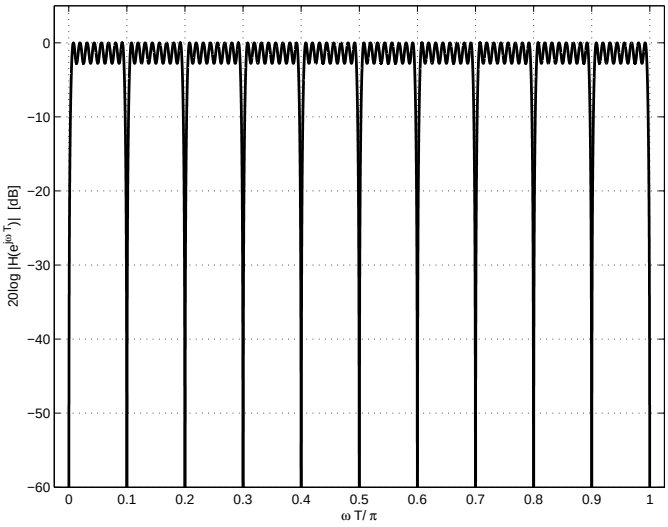


Fig. 4. Amplitude frequency response of the comb FIR filter of type 2.

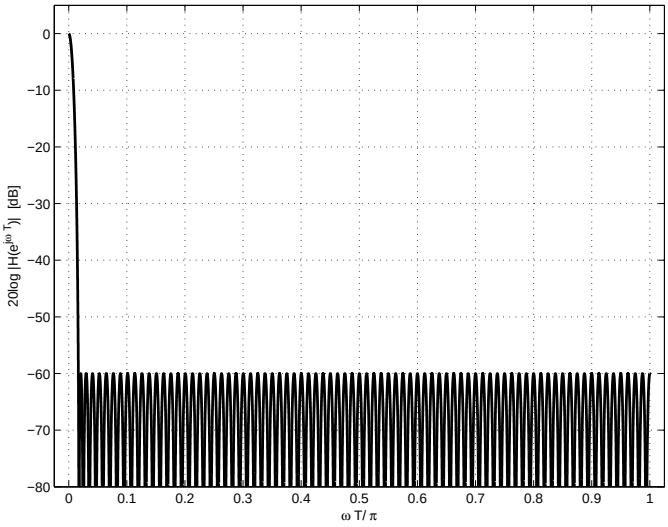


Fig. 5. Amplitude frequency response of the DC-pass FIR filter.

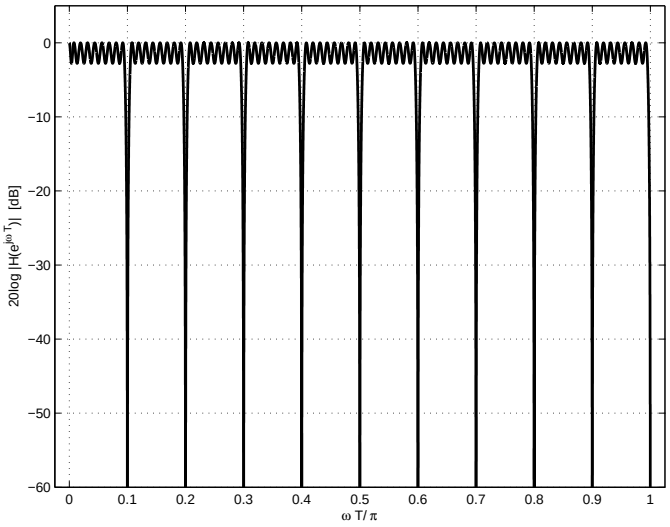


Fig. 6. Amplitude frequency response of the comb5 FIR filter.

response $h(k)$ of the length 791 coefficients of the comb5 FIR filter is calculated using (15). The amplitude frequency response $20\log|H(e^{j\omega T})|$ [dB] of the comb5 FIR filter is shown in Fig. 6.

VII. Conclusion

In this paper a fifth type of the comb FIR filter was introduced and its design procedure was presented. The coefficients of the impulse response of the filter are evaluated based on the filter specification. Independent control can be exercised over the number of notch bands, the width of the notch bands and attenuations in both pass bands and stop bands.

Acknowledgement

This activity was supported by the grant No. MSM6840770014, Ministry of Education, Czech Republic.

References

[1] Sanjit K. Mitra, Digital Signal Processing A Computer-Based Approach, McGraw-Hill, 1998.

[2] P. Zahradnik, M. Vlček, Analytical Design Method for Optimal Equiripple Comb FIR Filters, *IEEE Transactions on Circuits and Systems - II*, Vol. 52, No. 2, February 2005, pp. 112-115.

[3] P. Zahradnik, M. Vlček, R. Unbehauen, P. Design of Optimal Comb FIR Filters-Speed and Robustness, *IEEE Signal Processing Letters*, Vol. 16, No. 6, 2009, pp. 465-468.

[4] P. Zahradnik, M. Vlček, Note on the Design of an Equiripple DC-Notch FIR Filter, *IEEE Transactions on Circuits and Systems - II*, Vol. 54, No. 2, February 2007, pp. 196-199.

 ISBN:978-1-61208-113-7

Outage Performance Analysis of Alamouti STBC in Backward Link for Wireless Cooperative Networks

Wooju Lee, Dongweon Yoon, Seounghun Jee, and Jaeyoon Lee
Dept. of Electronics and Communications Eng.
Hanyang University
Seoul, Korea
E-mail: dwyoon@hanyang.ac.kr

Zhengyuan Xu
Dept. of Electrical Engineering
University of California Riverside
CA, USA

***Abstract*—In this paper, we propose a cooperative diversity scheme to improve the end-to-end outage performance via Alamouti space-time block coding with two transmit antennas at a source node in a backward link and a best-relay selection scheme in a forward link. We derive an exact closed-form expression of the outage probability for the proposed scheme over a Rayleigh fading channel and analyze end-to-end outage probabilities.**

Keywords-cooperative communication; decode-and-forward; selective relaying; space-time block coding; outage probability.

I. INTRODUCTION

Cooperative diversity has received considerable attention in recent years to increase the reliability of wireless networks. This technique uses one or more relay nodes to forward signals transmitted from the source node to the destination node [1]. Various schemes have been proposed to achieve cooperative diversity, such as amplify-and-forward (AF) and/or decode-and-forward (DF) schemes [2]-[8]. In this paper, we focus on the DF relaying, in which each relay node fully decodes, re-encodes, and then retransmits the source signals.

Approximations of the outage probabilities for a single-relay case were provided in [2] for a high signal-to-noise ratio (SNR) and in [3] for a low SNR. For a multiple-relay case, approximate and exact expressions of the outage probabilities based on distributed space-time coding cooperative schemes [2] were presented in [4], [5] and [6]. A best-relay selection scheme (BRSS), which has a simple cooperative diversity scheme in a multiple-relay scenario, was proposed in [7], where the diversity-multiplexing tradeoff of the BRSS was identical to that of the more complex distributed space-time coding cooperative schemes presented in [2]. For the BRSS, exact expression of the outage probability was provided in [8].

In this paper, we propose a cooperative diversity scheme to improve the end-to-end outage performance via Alamouti space-time block coding (STBC) with two transmit antennas at a source node in the backward link and the BRSS in a forward link from the relay nodes to the destination node. We derive an exact closed-form expression of the outage probability for the proposed scheme over a Rayleigh fading channel and analyze the end-to-end outage probability compared to the BRSS.

The paper is organized as follows: In Section II, we present the outage probability for the proposed scheme. Numerical results are presented in Section III, and the final section gives the conclusions.

II. OUTAGE PROBABILITY

We consider a half-duplex dual-hop communication scenario, where direct path is assumed to be blocked or to have poor connection due to an intermediate wall. In this case, the communication between a source node S and a destination node D is only possible via M relay nodes. We also assume that the channel is a Rayleigh fading channel with coherence time long enough for the system to complete transmitting a block of data.

The BRSS transmission between the source node S and destination node D is established during two time slots. In the first time slot, the source node S transmits the data to the M relay nodes. After the M relay nodes decode the data from the source node S, the selected relay node k among all the M relay nodes forwards the re-encoded data to the destination node D in the second time slot.

The mutual information of the BRSS between the source node S and the relay node k, I_{BRSS}^{Sk}, is given by [8]

$$I_{BRSS}^{Sk} = \frac{1}{2}\log_2\left(1+\frac{|h_{Sk}|^2 P}{N_0 W}\right) \quad (1)$$

where $P \triangleq E\left\{|s|^2\right\}$ is the average power of a transmit signal s, h_{Sk} is the complex channel gain between the source node S and the relay node k, N_0 is the noise power spectral density, and W is the transmission bandwidth. Similarly, the mutual information of the BRSS between the relay node k and the destination node D, I_{BRSS}^{kD}, is given by [8]

$$I_{BRSS}^{kD} = \frac{1}{2}\log_2\left(1+\frac{|h_{kD}|^2 P}{N_0 W}\right) \quad (2)$$

where h_{kD} is the complex channel gain between the relay node k and the destination node D. In (1) and (2), $|h_{Sk}|^2 P / N_0 W$ is the instantaneous SNR between the source node S and the relay node k and $|h_{kD}|^2 P / N_0 W$ is the instantaneous SNR between the relay node k and the destination node D. Note that, the factors of 1/2 in (1) and (2) account for the fact that the transmissions occur over two time slots. Then, the maximum instantaneous end-to-end mutual information of the BRSS is [8]

 ISBN:978-1-61208-113-7

$$I_{BRSS} = \max_{k \in K} \min\left(I_{BRSS}^{Sk}, I_{BRSS}^{kD}\right) \quad (3)$$

where K = {1,…, k,…, M}. The relay node that leads to the maximum in (3) is designated as the selected relay node k.

The outage probability of the BRSS can be written as [8]

$$P_{out}^{BRSS} = \prod_{k=1}^{M}\left[1 - \exp\left(-\left(\lambda_{Sk} + \lambda_{kD}\right)\left(2^{2R} - 1\right)\frac{N_0 W}{P}\right)\right] \quad (4)$$

where R is the fixed threshold rate, λ_{Sk} and λ_{kD} denote the channel conditions of the backward and forward links, respectively, which are the reciprocals of the expected values for the exponential random variables $x_{Sk} = |h_{Sk}|^2$ and $x_{kD} = |h_{kD}|^2$. Since channel gain h can be modeled as a complex Gaussian random variable, it is straightforwardly to show that the probability density function (PDF) of $x = |h|^2$ becomes the exponential PDF $f_X(x) = \lambda \exp(-\lambda x)$, $x \geq 0$, with parameter λ.

In order to improve the end-to-end outage performance, we propose a cooperative diversity scheme via Alamouti STBC with two transmit antennas at the source node S in the backward link and the BRSS in the forward link. The proposed scheme can obtain an additional transmit diversity gain in the backward link from two transmit antennas at the source node S. The proposed communication between the source node S and the destination node D is performed during four time slots without loss of rate.

In the first two time slots, the source node S transmits the two space-time block coded symbols to the M relay nodes. After the M relay nodes decode the two space-time block coded symbols from the source node S, the selected relay node k among all the M relay nodes forwards the two re-encoded symbols to the destination node D in the remaining two time slots. Then, the maximum instantaneous end-to-end mutual information of the proposed scheme is

$$I_{Prop.} = \max_{k \in K} \min\left(I_{Prop.}^{Sk}, I_{Prop.}^{kD}\right) \quad (5)$$

where $I_{Prop.}^{Sk}$ is the mutual information between the source node S and the relay node k, $I_{Prop.}^{kD}$ is the mutual information between the relay node k and the destination node D. The mutual information between the source node S and the relay node k is given by

$$I_{Prop.}^{Sk} = \frac{1}{2}\log_2\left(1 + \sum_{i=0}^{1}\frac{|h_{Sk,i}|^2 P}{2N_0 W}\right) \quad (6)$$

where $h_{Sk,i}$ is the complex channel gain between the source node S and the relay node k obtained by using the i-th transmit antenna in the source node S. Note that, the factor of $(P/2)$ in (6) accounts for the fact that the total transmission power is the same as that of the BRSS. The mutual information between the relay node k and the destination node D is the same as (2) and is given by

$$I_{Prop.}^{kD} = \frac{1}{2}\log_2\left(1 + \frac{|h_{kD}|^2 P}{N_0 W}\right). \quad (7)$$

The outage probability of the proposed scheme to maximize the minimum mutual information between the backward and forward links can be expressed as

$$\begin{aligned} P_{out}^{PROPOSED} &= P\left\{\max_{k \in K} \min\left(I_{Prop.}^{Sk}, I_{Prop.}^{kD}\right) < R\right\} \\ &= \prod_{k=1}^{M} P\left\{\min\left(I_{Prop.}^{Sk}, I_{Prop.}^{kD}\right) < R\right\} \\ &= \prod_{k=1}^{M}\left[1 - \left\{1 - P\left(I_{Prop.}^{Sk} < R\right)\right\}\left\{1 - P\left(I_{Prop.}^{kD} < R\right)\right\}\right] \end{aligned} \quad (8)$$

where $P\left(I_{Prop.}^{kD} < R\right)$ is the outage probability between the relay node k and the destination node D and $P\left(I_{Prop.}^{Sk} < R\right)$ is the outage probability between the source node S and the relay node k. We can obtain the outage probabilities, respectively, by

$$P\left(I_{Prop.}^{kD} < R\right) = 1 - \exp\left(-\lambda_{kD}\left(2^{2R} - 1\right)\frac{N_0 W}{P}\right) \quad (9)$$

and

$$P\left(I_{Prop.}^{Sk} < R\right) = P\left(\sum_{i=0}^{1} x_{Sk,i} < \left((2^{2R} - 1)\frac{2N_0 W}{P}\right)\right) \quad (10)$$

where λ_{kD} and $\lambda_{Sk,i}$ are the parameters of the exponential random variables $x_{kD} = |h_{kD}|^2$ and $x_{Sk,i} = |h_{Sk,i}|^2$, respectively. Note that the outage probability between the source node S and the relay node k in (10) is simply the cumulative distribution function (CDF) of the sum, $x_{Sk,sum} = \sum_{i=0}^{1} x_{Sk,i}$. We assume that the channels between two transmit antennas at the source node S and each relay k in the backward link are independent and identically distributed $\left(\lambda_{Sk} = \lambda_{Sk,0} = \lambda_{Sk,1}\right)$. Under this assumption, we use the method of moment generating function (MGF) to find the CDF of $x_{Sk,sum}$. The MGF of $x_{Sk,sum}$ can be obtained as [9]

$$M_{Sk,sum}(s) = \left(\lambda_{Sk} / \left(s + \lambda_{Sk}\right)\right)^2. \quad (11)$$

Then, applying inverse Laplace transform and integration for (11), we can obtain the outage probability in the backward link as

$$P\left(I_{Prop.}^{Sk} < R\right) = 1 - \left(1 + 2\lambda_{Sk}\Psi\right)\exp\left(-2\lambda_{Sk}\Psi\right) \quad (12)$$

where $\Psi = (2^{2R} - 1)N_0 W / P$.

Finally, by using (8), (9), and (12), a closed-form expression of the end-to-end outage probability for the proposed scheme can be obtained as

$$P_{out}^{PROPOSED} = \prod_{k=1}^{M}\left[1 - \left(1 + 2\lambda_{Sk}\Psi\right)\exp\left(-\left(\lambda_{kD} + 2\lambda_{Sk}\right)\Psi\right)\right]. \quad (13)$$

III. NUMERICAL RESULTS

We compare the outage probabilities of the proposed scheme and the BRSS. In this paper, we assume that λ_{Sk} and λ_{kD} are uniformly distributed in [0, 1], and the target threshold rate R is 1 bit/sec/Hz.

 ISBN:978-1-61208-113-7

In Figure 1, we illustrate the outage probabilities of the proposed scheme and the BRSS, when M = 2, 4, and 10, respectively. This figure shows that the proposed scheme outperforms the BRSS on the outage probabilities. When M = 2, 4, and 10 relays are employed under the assumed channel condition, the proposed scheme outperforms the BRSS at the outage probability of 10^{-4} by 2.95 dB, 2.38 dB, 1.13 dB, respectively. This is because the proposed scheme can obtain the additional transmit diversity gain in the backward link from two transmit antennas at the source node S.

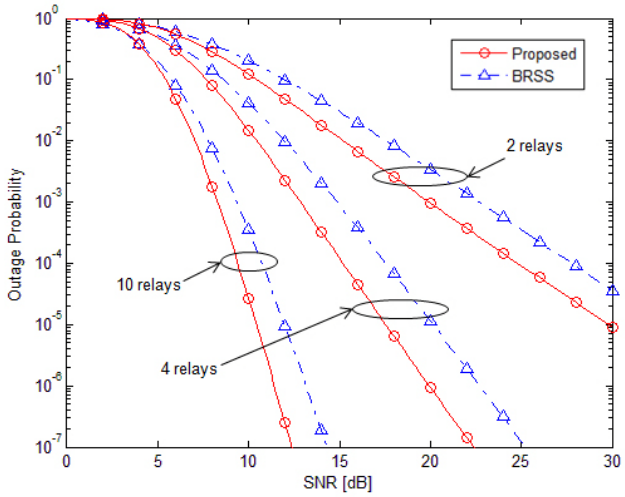


Figure 1. Outage probabilities of proposed scheme and BRSS with R = 1 bit/sec/Hz.

IV. CONCLUSIONS

In this paper, we proposed a cooperative diversity scheme with DF relaying to improve the end-to-end outage performance via Alamouti STBC with two transmit antennas at the source node. We derived an exact closed-form expression of the outage probability for the proposed scheme over a Rayleigh fading channel and analyzed the end-to-end probability compared to the BRSS. From the results, it was confirmed that end-to-end outage probability of the proposed scheme outperformed the BRSS in wireless cooperative networks.

ACKNOWLEDGMENT

This research was supported by NSL program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2010-0015083).

REFERENCES

[1] A. Sendonaris, E. Erkip, and B. Aazhang, “User cooperation diversity–part I: system description,” *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[2] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, “Cooperative diversity in wireless networks: Efficient protocols and outage behavior,” *IEEE Trans. Infom. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[3] A. S. Avestimehr and D. N. C. Tse, “Outage capacity of the fading relay channel in the low-SNR regime,” *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1401–1415, Apr. 2007.

[4] Y. Zhao, R. Adve, and T. J. Lim, “Outage probability at arbitrary SNR with cooperative diversity,” *IEEE Commun. Lett.*, vol. 9, no. 8, pp. 700–702, Aug. 2005.

[5] N. C. Beaulieu and J. Hu, “A closed-form expression for the outage probability of decode-and-forward relaying in dissimilar Rayleigh fading channels,” *IEEE Commun. Lett.*, vol. 12, pp. 813–815, Dec. 2006.

[6] J. Hu and N. C. Beaulieu, “Performance analysis of decode-and-forward relaying with selection combining,” *IEEE Commun. Lett.*, vol. 11, no. 6, pp. 489–491, Jun. 2007.

[7] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, “A simple cooperative diversity method based on network path selection,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 659–672, Mar. 2006.

[8] K. Woradit, T.Q.S. Quek, W. Suwansantisuk, H.A. Wymeersch, L. Wuttisittikulkij, and M.Z. Win, “Outage behavior of selective relaying schemes,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 3890–3895, Aug. 2009.

[9] J.G. Proakis, *Digital Communications*, McGraw-Hill, 1995.

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

 ISBN:978-1-61208-113-7

Exact Error Probabilities Analysis of Arbitrary 2-D Modulation-OFDM Systems with I/Q Imbalances in Frequency-Flat Rayleigh Fading Channel

Jaeyoon Lee, Dongweon Yoon, Kyongkuk Cho, and Wooju Lee
Department of Electronics and Computer Engineering
Hanyang University, Seoul, Korea
Email: jylee1988@gmail.com, dwyoon@hanyang.ac.kr, kyongkuk@gmail.com, and universe@hanyang.ac.kr

***Abstract*—In OFDM systems, in-phase and quadrature (I/Q) imbalances generated in the analog front-end introduce inter-channel interference and, consequently, error performance degradation. This paper provides an exact expression involving the two-dimensional (2-D) Gaussian Q-function for the error probability of an arbitrary 2-D modulated OFDM signal with I/Q imbalances in frequency-flat Rayleigh fading channel.**

***Index Terms*—I/Q imbalance, Probability of error, Inter-channel interference**

I. INTRODUCTION

In-phase and quadrature (I/Q) amplitude and phase imbalances are inevitably caused by signal processing in the analog components such as I/Q mixers, phase shifters, filters, and analog/digital converters within I/Q branches. In the implementation of a modern wireless communication system, I/Q imbalances act as one of the main impairments degrading system performance. Particularly in the orthogonal frequency division multiplexing (OFDM) schemes adapted in a number of wireless communication systems such as DAB, DVB-T, WLAN (802.11 a/g/n), WPAN (802.15.3a), WMAN (802.16 a/d/e), and MBWA (802.20), I/Q imbalances introduce inter-channel interference (ICI) and nonlinearly distort the baseband signals [1], [2]. The effects of I/Q imbalances on OFDM system performance have been analyzed by computer simulations, and several compensation techniques have been reported in many places in the literature [3]-[6].

In general, for a single-carrier system, I/Q imbalances in the receiver lead to a correlation between I/Q branches, and the correlation is revealed as the variations of the received signal points and the noise distribution on the constellation [7], [8]. For a single-carrier system, a method was recently provided to exactly analyze the effect of I/Q imbalances generated in the analog front-end of the receiver on the error performance [9]. In this paper, we derive an exact expression for the error probability of an arbitrary 2-D modulated signal with I/Q imbalances in an OFDM system over frequency-flat Rayleigh fading channel. Through a computer simulation, we verify the validity of the result obtained from the derived expression.

II. SYSTEM MODEL

Fig. 1 shows a typical OFDM transceiver where we assume, for simplicity of analysis, I/Q amplitude and phase imbalances take place in the receiver side. As shown in Fig. 1, the OFDM transmitter undergoes an ideal complex up-conversion, but the received signals are affected by I/Q imbalances in the analog front-end of the receiver.

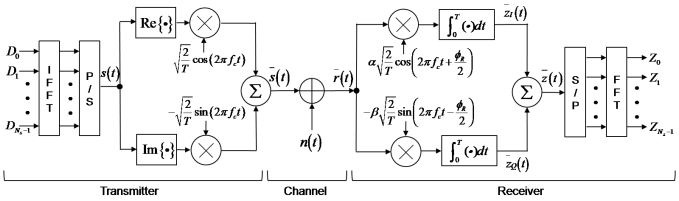


Fig. 1. OFDM transmitter and receiver with I/Q imbalances.

Assume that $D_k \in \mathbf{t\_s} = [t\_s_1\ t\_s_2\ ...\ t\_s_M]$ is the complex symbol modulated by an arbitrary 2-D M-ary modulator, and is transmitted on subcarrier k in the OFDM system, where $\mathbf{t\_s}$ is a set of M signal points transmitted through the OFDM system. Then, D_k can be defined as

$$D_k = \zeta_{i,k}\sqrt{E_s}e^{j\psi_{i,k}}, k = 0, 1, ..., N_s - 1, i = 1, 2, ..., M \qquad (1)$$

where $\psi_{i,k}$ and $\zeta_{i,k}\sqrt{E_s}$ are the phase and the amplitude of the i-th signal point transmitted on subcarrier k, E_s is the average symbol energy, $\zeta_{i,k}$ is a scale factor which varies with the position of the signal point, and N_s is the number of subcarriers.

The ideally up-converted transmitted signal at the transmitter is formally given by

$$\overline{s}(t) = \frac{1}{\sqrt{2T_s}}\left(s(t)\,e^{j2\pi f_c t} + s^*(t)\,e^{-j2\pi f_c t}\right) \qquad (2)$$

where $s(t) = \sum_{k=0}^{N_s-1} D_k e^{j2\pi kt/T_s}$ is a baseband OFDM signal and $s^*(t)$ denotes the complex conjugate of $s(t)$.

After the received bandpass signal, $\overline{r}(t) = \overline{s}(t) + n(t)$, is down-converted by mixing with $x_{LO}^{unbalanced}(t)$, the baseband signal, $\overline{z}(t)$, distorted by I/Q imbalances in time domain, is

expressed as

$$\begin{aligned}\overline{z}(t) &= \overline{z}_I(t) + j\,\overline{z}_Q(t) \\ &= K_1 s(t) + K_2 s^*(t) + \sqrt{\tfrac{2}{T_s}}\cdot \\ &\quad \left(\int_0^T K_2 n(t)\, e^{j2\pi f_c t} dt + \int_0^T K_1 n(t)\, e^{-j2\pi f_c t} dt\right) \\ &= K_1 s(t) + K_2 s^*(t) + n_I + j n_Q\end{aligned} \tag{3}$$

where $K_1 = (\alpha e^{-j\frac{\varphi_R}{2}} + \beta e^{j\frac{\varphi_R}{2}})/2$ and $K_2 = (\alpha e^{j\frac{\varphi_R}{2}} - \beta e^{-j\frac{\varphi_R}{2}})/2$ are imbalance coefficients. In (4), $n(t)$ is an additive Gaussian noise with zero mean and variance of σ^2; n_I and n_Q are noise components on I/Q branches, and can be expressed as

$$\begin{aligned} n_I &= \tfrac{1}{\sqrt{T_s}}\alpha \int_0^T n(t)\sqrt{2}\cos\left(2\pi f_c t + \tfrac{\varphi_R}{2}\right) dt \\ n_Q &= \tfrac{1}{\sqrt{T_s}}\beta \int_0^T n(t)\left(-\sqrt{2}\sin\left(2\pi f_c t - \tfrac{\varphi_R}{2}\right)\right) dt \end{aligned} \tag{4}$$

where α and β are the amplitude gains on the I/Q branches, which represent the amplitude imbalances, and φ_R is the deviation from the perfect phase quadrature, which represents the phase imbalance. We assume the condition of $\alpha^2+\beta^2 = 2$ to leave the signal power unchanged, and define gain ratio, γ, and amplitude imbalance, ε, as follows [10]:

$$\gamma = \alpha/\beta\ , \quad \varepsilon = \gamma - 1. \tag{5}$$

Note that n_I and n_Q have joint Gaussian distribution with zero mean, $E[n_I^2] = \alpha^2\sigma^2$, $E[n_Q^2] = \beta^2\sigma^2$, and $E[n_I n_Q] = \rho_{IQ}\sigma^2$, where $\rho_{IQ} = \alpha\beta\sin\varphi_R$ denotes the correlation coefficient between I/Q axes.

A noise distribution has an elliptical shape before FFT, which means that noise components on the I/Q axes are correlated. After FFT, the complex symbol Z_k, passed through the FFT block, can be expressed as

$$\begin{aligned} Z_k &= FFT\left[K_1 s(t) + K_2 s^*(t) + n_I + j n_Q\right] \\ &= K_1 D_k + K_2 D^*_{N_s-k} + N_{I\_k} + j N_{Q\_k}, \\ &\qquad k = 0,\, 1,\, ...,\, N_s - 1 \end{aligned} \tag{6}$$

where N_{k_I} and N_{k_Q} are the noise components of subcarrier k on I/Q branches, which have joint Gaussian distribution with zero mean, $E[N_{k\_I}^2] = E[N_{k\_Q}^2] = \sigma^2$, and $E[N_{k\_I} N_{k\_Q}] = 0$. Note that noise components on the I/Q axes become uncorrelated after FFT. From (6), we also note that D_{N_s-k} causes ICI to D_k [5].

Substituting (1), K_1 and K_2 into (6), we can obtain

$$\begin{aligned} Z_k &= S_{k\_I} + j S_{k\_Q} + N_{k\_I} + j N_{k\_Q}, \\ &\qquad k = 0,\, 1,\, \ldots,\, N_s - 1 \end{aligned} \tag{7}$$

where S_{k_I} and S_{k_Q} are the received signal components on I/Q axes, expressible as

$$\begin{aligned} &S_{k\_I} \\ &= \tfrac{\zeta_{i,\,k}\sqrt{E_s}}{2}\left(\alpha\cos\left(\psi_{i,\,k} - \tfrac{\varphi_R}{2}\right) + \beta\cos\left(\psi_{i,\,k} + \tfrac{\varphi_R}{2}\right)\right) \\ &\quad + \tfrac{\zeta_{m,\,N_s-k}\sqrt{E_s}}{2}\left(\alpha\cos\left(\psi_{m,\,N_s-k} - \tfrac{\varphi_R}{2}\right)\right. \\ &\qquad \left. -\beta\cos\left(\psi_{m,\,N_s-k} + \tfrac{\varphi_R}{2}\right)\right) \\ &S_{k\_Q} \\ &= \tfrac{\zeta_{i,\,k}\sqrt{E_s}}{2}\left(\alpha\sin\left(\psi_{i,\,k} - \tfrac{\varphi_R}{2}\right) + \beta\sin\left(\psi_{i,\,k} + \tfrac{\varphi_R}{2}\right)\right) \\ &\quad + \tfrac{\zeta_{m,\,N_s-k}\sqrt{E_s}}{2}\left(-\alpha\sin\left(\psi_{m,\,N_s-k} - \tfrac{\varphi_R}{2}\right)\right. \\ &\qquad \left. +\beta\sin\left(\psi_{m,\,N_s-k} + \tfrac{\varphi_R}{2}\right)\right), \\ &\qquad i = m\ = 1,\, 2,\, ...,\, M \end{aligned} \tag{8}$$

where ψ_{m,N_s-k} and $\zeta_{m,N_s-k}\sqrt{E_s}$ are the phase and the amplitude of the m-th signal point on subcarrier $(N_s - k)$, respectively. Note that the second terms of S_{k_I} and S_{k_Q} induce the ICI, that is, the complex symbol transmitted on subcarrier $(N_s - k)$ interferes with the complex symbol transmitted on subcarrier k.

III. Exact Error Probability Expression for an Arbitrary 2-D Modulated OFDM Signal with I/Q Imbalances

In this section, we derive the exact SER/BER expressions for an arbitrary 2-D modulated OFDM signal with I/Q imbalances in AWGN, frequency flat Rayleigh fading, and frequency selective Rayleigh fading channels, respectively.

A. AWGN Channel

Fig. 2 shows the geometry of the correct decision region R_{s_1} for the received signal, r_s_1, when the signal transmitted on subcarrier k is t_s_1. This geometry acts as a basic shape for the evaluation of exact error probability for I/Q unbalanced case [9]. As shown in Fig. 2, the distortion due to ICI which arises from I/Q imbalances results in a shift of the received signal point on the constellation. In Fig. 2, $S_k|t\_s_i$ denotes the transmitted signal on subcarrier k interfered by the signal t_s_i transmitted on subcarrier $(N_s - k)$. I/Q noise components $(N_{k\_I},\, N_{k\_Q})$ added to the transmitted signal also change the position of the received signal points, $r\_s_i,\ \ i = 1,\, 2,\, ...,\, M$.

To obtain the conditional probability $P_k\{r\_s_1 \in R_{s_1} | S_t = t\_s_1\}$ that the signal point r_s_1 received through subcarrier k falls into a shaded region $R_{s_1} = X_1^- C X_2^+$, which represents the correct region for t_s_1 in Fig. 2, we use the coordinate rotation and shifting technique well explained in [9] as follows:

$$\begin{bmatrix} X_j \\ Y_j + d_j \end{bmatrix} = \begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{bmatrix} \begin{bmatrix} I \\ Q \end{bmatrix}, \quad j = 1,\, 2 \tag{9}$$

where d_j, $j = 1,\, 2$ are the distances from the origin to the X_j, $j = 1,\, 2$ axes which denote the decision boundaries; I and Q have joint Gaussian distribution with $E[I] = S_{k\_I}$, $E[Q] = S_{k\_Q}$, $Var[I] = Var[Q] = \sigma^2$, and $COV[I\,Q] = 0$. After rotational transformation (9), Y_1 and Y_2 which are axes newly made by the coordinate rotations with $-\theta_1$ and θ_1 from Q axis have joint Gaussian probability density function (pdf)

 ISBN:978-1-61208-113-7

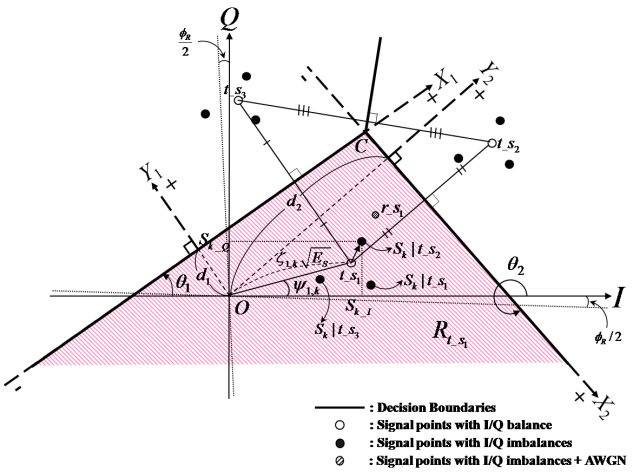


Fig. 2. Decision region and boundaries for ternary signal points.

$f(y_1, y_2, \rho_{Y_1Y_2})$ with

$$\begin{cases} E[Y_j] = S_{k\_Q}\cos\theta_j - S_{k\_I}\sin\theta_j - d_j, \quad j = 1, 2 \\ Var\,[Y_j] = \sigma^2 \\ \rho_{Y_1Y_2} = \cos\left(\theta_1 - \theta_2\right) \end{cases} \tag{10}$$

where $Var[Y_j]$ is the variance of Y_i; $\rho_{Y_1Y_2}$ is the correlation coefficient between Y_1 and Y_2.

Consequently, the conditional probability $P_k\{r\_s_1 \in R_{t\_s_1}|S_t = t\_s_1\}$ in Fig. 2 can be obtained as

$$\begin{aligned} &P_k\left\{r\_s_1 \in R_{t\_s_1}|S_t = t\_s_1\right\} = P_k\left\{Y_1 \le 0,\, Y_2 \le 0\right\} \\ &= \int_{-\infty}^{0}\int_{-\infty}^{0} f\left(y_1, y_2, \rho_{Y_1Y_2}\right)dy_2dy_1 \\ &= \int_{-\infty}^{-\frac{E[Y_1]}{\sqrt{Var[Y_1]}}} \int_{-\infty}^{-\frac{E[Y_2]}{\sqrt{Var[Y_2]}}} \left[2\pi\sqrt{1-\rho_{Y_1Y_2}^2}\right]^{-1} \cdot \\ &\qquad \exp\left[-\frac{1}{2}\left(\frac{u^2 - 2\rho_{Y_1Y_2}uv + v^2}{1-\rho_{Y_1Y_2}^2}\right)\right] dvdu \\ &= Q\left(\frac{E[Y_1]}{\sqrt{Var[Y_1]}},\, \frac{E[Y_2]}{\sqrt{Var[Y_2]}}; \rho_{Y_1Y_2}\right). \end{aligned} \tag{11}$$

Finally, since the baseband signal transmitted on subcarrier $(N_s - k)$, which causes ICI to the signal transmitted on subcarrier k in the receiver, is one of the M signals, $t\_s_i,\ i = 1, 2, ..., M$, the average SER for a signal point t_s_1 transmitted on subcarrier k can be written as

$$\begin{aligned} &P_{ser\_t\_s_1} \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\left(1 - P_k\left\{r\_s_1 \in R_{t\_s_1}|S_t = t\_s_1\right\}\right)P\{t\_s_1\}\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\left(1 - Q\left(\frac{E[Y_1]}{\sqrt{Var[Y_1]}}, \frac{E[Y_2]}{\sqrt{Var[Y_2]}}; \rho_{Y_1Y_2}\right)\right)P\{t\_s_1\}\right] \end{aligned} \tag{12}$$

where $P\{t\_s_1\}$ is *a priori* probability for the transmitted signal point, and m denotes an index of the signal point transmitted on subcarrier $(N_s - k)$ which causes interference, as in (8).

In general, the decision region of a transmitted signal point is a polygon that may be either closed or open [12], and the decision region can be expressed as a linear combination of the basic shapes [9], [13]. Therefore, an exact error probability expression for the signal point with the polygonal decision region can be obtained by using the probability of (12).

When a signal point t_s_1 that has a closed region case with u-sided polygonal shape is transmitted, the SER for the signal point is derived as

$$\begin{aligned} &P^c_{ser\_t\_s_1} \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\left(1 - P_k\left\{r\_s_1 \in R_{t\_s_1^c}|S_t = t\_s_1^c\right\}\right)P\left\{t\_s_1^c\right\}\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\left\{1 - \left(Q\left(\frac{E[Y_1]}{\sqrt{Var[Y_1]}}, \frac{E[Y_2]}{\sqrt{Var[Y_2]}}; \rho_{Y_1Y_2}\right)\right.\right.\right. \\ &\qquad \left. + Q\left(\frac{E[Y_1]}{\sqrt{Var[Y_1]}}, \frac{E[Y_4]}{\sqrt{Var[Y_4]}}; \rho_{Y_1Y_4}\right)\right. \\ &\quad \left.\left. - \sum_{i=2}^{u-1} Q\left(\frac{E[Y_i]}{\sqrt{Var[Y_i]}}, -\frac{E[Y_{i+1}]}{\sqrt{Var[Y_{i+1}]}}; -\rho_{Y_iY_{i+1}}\right)\right)\right\}P\left\{t\_s_1^c\right\}\right]. \end{aligned} \tag{13}$$

And, when a signal point t_s_1 that has a open region case with v-sided polygonal shape is transmitted, the SER for the signal point is derived as

$$\begin{aligned} &P^o_{ser\_t\_s_1} \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\left(1 - P_k\left\{r\_s_1 \in R_{t\_s_1^o}|S_t = t\_s_1^o\right\}\right)P\left\{t\_s_1^o\right\}\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\left\{1 - \left(Q\left(\frac{E[Y_1]}{\sqrt{Var[Y_1]}}, \frac{E[Y_2]}{\sqrt{Var[Y_2]}}; \rho_{Y_1Y_2}\right)\right.\right.\right. \\ &\quad \left.\left. - \sum_{i=2}^{v-1} Q\left(\frac{E[Y_i]}{\sqrt{Var[Y_i]}}, -\frac{E[Y_{i+1}]}{\sqrt{Var[Y_{i+1}]}}; -\rho_{Y_iY_{i+1}}\right)\right)\right\}P\left\{t\_s_1^o\right\}\right]. \end{aligned} \tag{14}$$

Finally, the average SER of an arbitrary 2-D modulated OFDM signal with I/Q imbalances on k-th subcarrier is obtained as

$$\begin{aligned} &P_k^{ser} \\ &= \sum_{i=1}^{U}\left[\frac{1}{M}\sum_{m=1}^{M}\left[\left(1 - P_k\left\{r\_s_i^c \in R_{s_i}^c|S_t = t\_s_i^c\right\}\right)P\{t\_s_i^c\}\right]\right] \\ &+ \sum_{i=1}^{V}\left[\frac{1}{M}\sum_{m=1}^{M}\left[\left(1 - P_k\left\{r\_s_i^o \in R_{s_i}^o|S_t = t\_s_i^o\right\}\right)P\{t\_s_i^o\}\right]\right] \end{aligned} \tag{15}$$

where U is the number of signal points with the closed correct region, V is the number of signal points with the open correct region, and $M = U + V$.

Obtaining the exact BER expression for an arbitrary 2-D modulated OFDM signal with I/Q imbalances on k-th subcarrier is very tedious work, but exact BER performance is obtained by using the result of [14] in the form

$$\begin{aligned} P_k^{ber} = \frac{1}{\log_2 M}\sum_{l=1}^{M}\sum_{\substack{h=1 \\ h\neq l}}^{M}\left[\frac{1}{M}\sum_{m=1}^{M}\left[H\_d\left(t\_s_l,\, t\_s_h\right)\cdot\right.\right. \\ \left.\left. P_k\left\{r\_s_l \in R_{s_h}|S_t = t\_s_l\right\}\cdot P\left\{t\_s_l\right\}\right]\right] \end{aligned} \tag{16}$$

where $t\_s_i,\ i = 1, 2, ..., M$ are the transmitted symbol, R_{s_h} represents the decision region for the symbol t_s_h , and $H\_d(t\_s_l,\, t\_s_h)$ denotes the Hamming distance between t_s_l and t_s_h .

B. Frequency-Flat Rayleigh Fading Channel

To obtain the exact error probability for the arbitrary 2-D modulated OFDM system in frequency-flat Rayleigh fading channel, we first should derive a closed-form solution

of $\int_0^\infty Q\left(\Delta_1\sqrt{\gamma}, \Delta_2\sqrt{\gamma}; -\rho\right) \cdot f_\gamma(\gamma)\, d\gamma$ where $f_\gamma(\gamma) = \exp(-\gamma/\bar{\gamma})/\bar{\gamma}$, $\gamma \geq 0$. In the 2-D Gaussian Q-function, $Q(x, y; \rho)$, of (15) and (16), the range of x and y is $-\infty < x, y < \infty$. Therefore, rewriting the 2-D Gaussian Q-function by using the Crag representation [15] and its alternative expression [16], $\int_0^\infty Q\left(\Delta_1\sqrt{\gamma}, \Delta_2\sqrt{\gamma}; -\rho\right) \cdot f_\gamma(\gamma)\, d\gamma$ is expressed as

$$\begin{aligned} &\int_0^\infty Q\left(\Delta_1\sqrt{\gamma}, \Delta_2\sqrt{\gamma}; -\rho\right) f_\gamma(\gamma)\, d\gamma \\ &= \tfrac{1}{2}\mathrm{sgn}\left[\mathrm{sgn}(\Delta_1) + \mathrm{sgn}(\Delta_1\Delta_2)\right] - \tfrac{1}{2}\mathrm{sgn}(\Delta_1) \\ &\quad + \sum_{i=1}^{2} \mathrm{sgn}(\Delta_i)\, \tfrac{1}{2\pi} \int_0^{\omega_i} \int_0^\infty \exp\left(-\tfrac{\Delta_i^2\gamma}{2\sin^2\theta}\right) f_\gamma(\gamma)\, d\gamma d\theta, \\ &\qquad -\infty < \Delta_i < \infty \text{ and } 0 \leq \omega_i \leq \pi \end{aligned} \tag{17}$$

where $\omega_1 = \pi/2 + \sin^{-1}((\rho\Delta_1 - \Delta_2)/\sqrt{(\Delta_1^2 - 2\rho\Delta_1\Delta_2 + \Delta_2^2)})$; $\omega_2 = \pi/2 + \sin^{-1}((\rho\Delta_2 - \Delta_1)/\sqrt{(\Delta_1^2 - 2\rho\Delta_1\Delta_2 + \Delta_2^2)})$. And then, applying the moment generating function (MGF) of γ, *i.e.* $M_\gamma(s) = (1 - \bar{\gamma}s)^{-1}$, to (17), we can straightforwardly obtain the following closed-form expression from [17, eq. (5A.24)] as follows:

$$\begin{aligned} &\int_0^\infty Q\left(\Delta_1\sqrt{\gamma}, \Delta_2\sqrt{\gamma}; -\rho\right) \cdot f_\gamma(\gamma)\, d\gamma \\ &= \tfrac{1}{2}\mathrm{sgn}\left[\mathrm{sgn}(\Delta_1) + \mathrm{sgn}(\Delta_1\Delta_2)\right] - \tfrac{1}{2}\mathrm{sgn}(\Delta_1) \\ &\quad + \tfrac{1}{2}\sum_{i=1}^{2} \mathrm{sgn}(\Delta_i)\left[\tfrac{\omega_i}{\pi} - \tfrac{1}{\pi}\beta_i\left(\tfrac{\pi}{2} + \tan^{-1}\alpha_i\right)\right] \end{aligned} \tag{18}$$

where $\beta_i = \sqrt{c_i/1 + c_i}\mathrm{sgn}\omega_i$, $\alpha_i = -\beta_i \cot\omega_i$, and $c_i = \Delta_i^2\bar{\gamma}$.

Finally, applying the result of (18) to (15) and (16), the exact symbol and bit error probabilities for an arbitrary 2-D modulation based OFDM system with I/Q imbalances over frequency-flat Rayleigh fading channel can be derived in exact closed-form expressions.

IV. Numerical Results and Conclusions

The 2-D modulation format considered in this section is (4+12)-APSK, which have been adopted as standard modulation techniques for the DVB-S2 system and space data system [18], [19] for their robust performance in nonlinear high power amplifiers (HPA). Because the SIR for the practical imbalance values is in the order of 20-30dB [11], we consider the imbalance values of $\varphi_R = 3°$, $\gamma = 1.1$ $(SIR \approx 25dB)$ and $\varphi_R = 5°$, $\gamma = 1.2$ $(SIR \approx 20dB)$ in this paper. Fig. 3 shows the SER and BER curves for (4+12)-APSK based OFDM system over frequency-flat Rayleigh fading channel. From the figure we can see excellent matches between the results obtained from our exact expressions and computer simulations. We can also verify that the gap of error rate increases as the effect of I/Q imbalances becomes greater.

In this paper, we have provided an exact closed-form expression involving the 2-D Gaussian Q-function for the error probabilities of an arbitrary 2-D modulated OFDM signal with I/Q imbalances in frequency-flat Rayleigh fading channel, and analyzed the effect of I/Q imbalances on error performance. The result can be readily applied to numerical evaluation for various cases of practical interest involving unbalanced I/Q modulation in OFDM systems.

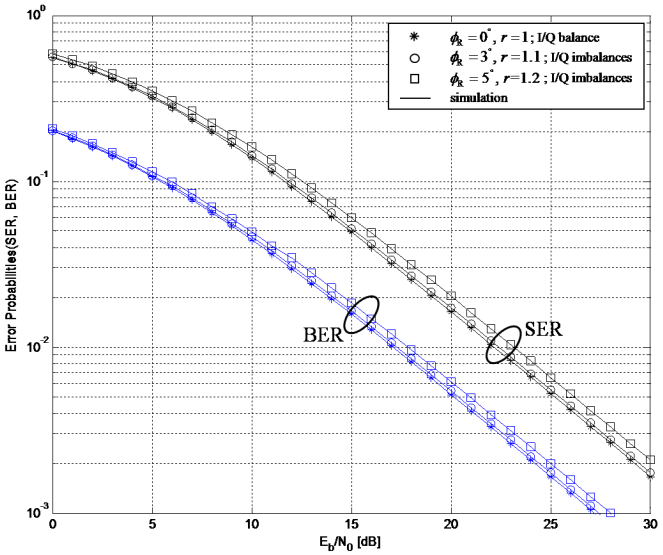


Fig. 3. SER and BER for (4+12)-APSK-based-OFDM system.

Acknowledgment

This research was supported by NSL program through the Korea Science and Engineering Foundation funded by the Ministry of Education, Science and Technology (2010-0015083).

References

[1] C. L. Liu, "Impacts of I/Q imbalance on QPSK-OFDM-QAM detection," *IEEE Trans. Consumer Electron.*, vol. 44, no. 3, pp. 984-989, Aug. 1998.

[2] M. Buchholz, A. Schuchert, and R. Hasholzner, "Effects of tuner IQ imbalance on multicarrier-modulation systems," in *Proc. IEEE ICCDCS 2000*, Cancun, Mar. 2000.

[3] M. Valkama, M. Renfors, and V. Koivunen, "Advanced methods for I/Q imbalance compensation in communication receivers," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2335-2344, Oct. 2001.

[4] A. Schuchert, R. Hasholzner, and P. Antoine, "A novel IQ imbalance compensation scheme for the reception of OFDM signals," *IEEE Trans. Consumer Electron.*, vol. 47, no. 3, pp. 313-318, Aug. 2001.

[5] A. Tarighat, R. Bagheri, and A. H. Sayed, "Compensation schemes and performance analysis of IQ imbalances in OFDM receivers," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3257-3268, Aug. 2005.

[6] A. Tarighat and A. H. Sayed, "MIMO OFDM receivers for systems with IQ imbalances," *IEEE Trans. Signal Process.*, vol. 53, no. 9, pp. 3583-3596, Sept. 2005.

[7] M. K. Simon and D. Divsalar, "Some new twists to problems involving the Gaussian probability integral," *IEEE Trans. Commun.*, vol. 46, no. 2, pp. 200-210, Feb. 1998.

[8] S. Park and D. Yoon, "An alternative expression for the symbol error probability of MPSK in the presence of I/Q unbalance," *IEEE Trans. Commun.*, vol. 52, issue 12, pp. 2079-2081, Dec. 2004.

[9] J. Lee, D. Yoon, and K. Hyun, "Exact and general expression for the error probability of arbitrary two-dimensional signaling with I/Q amplitude and phase unbalances," *IEICE Trans. Commun.*, vol. E89-B, no.12, pp. 3356-3362, Dec. 2006.

[10] J. K. Cavers and M. W. Liao, "Adaptive compensation for imbalance and offset losses in direct conversion transceivers," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 581-588, Nov. 1993.

[11] M. Valkama, M. Renfors, and V. Koivunen, "Blind signal estimation in conjugate signal models with application to I/Q imbalance compensation," *IEEE Signal Process. Lett.*, vol. 12, no. 11, pp. 733-736, Nov. 2005.

[12] L. Xiao and X. Dong, "The exact transition probability and bit error probability of two-dimensional signaling," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2600-2609, Sept. 2005.

 ISBN:978-1-61208-113-7

[13] L. Szczecinski, S. A?ssa, C. Gonzalez, and M. Bacic, "Exact evaluation of bit- and symbol-error rates for arbitrary 2-D modulation and nonuniform signaling in AWGN channel," *IEEE Trans. Commun.*, vol. 54, no. 6, pp. 1049-1056, June 2006.

[14] J. Lassing, E. G. Strom, E. Agrell, and T. Ottosson, "Computation of the exact bit-error rate of coherent M-ary PSK with Gray code bit mapping," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1758-1760, Nov. 2003.

[15] M. K. Simon, "A simpler form of the Craig representation for the two-dimensional joint Gaussian Q-function," *IEEE Commun. Lett.*, vol. 6, no. 2, pp. 49-51, Feb. 2002.

[16] S. Park and U J. Choi, "A generic Craig form for the two-dimensional Gaussian Q-function," *ETRI Journal*, vol. 29, no. 4, pp. 516-517, Aug. 2007.

[17] M. K. Simon and M. S. Alouini, *Digital Communication over Fading Channels*, 2nd ed, Wiley, 2005.

[18] "Digital Video Broadcasting; Second Generation Framing Structure, Channel Coding and Modulation Systems for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications," ETSI EN 302 307 v1.1.1.

[19] "Flexible serially concatenated convolutional turbo codes with near-Shannon bound performance for telemetry applications," Research and Development for Space Data System Standards, Experimental Specification CCSDS 131.2-O-1.

 ISBN:978-1-61208-113-7

Performance Issues in the Design of a VPN Resistant to Traffic Analysis

Claudio Ferretti, Alberto Leporati, Riccardo Melen
Dipartimento di Informatica, Sistemistica e Comunicazione
Università di Milano Bicocca
Milano, Italy
{ferretti, leporati, melen}@disco.unimib.it

Abstract— **This paper describes the architecture of BlankNet, a secure Virtual Private Network, which is capable of protecting its users against most kinds of attacks to traffic secrecy, including traffic analysis attacks. A key aspect of the BlankNet architecture is the definition of the allowed packet communication patterns: we call this pattern a *virtual topology*. Among the various possible solutions, the binary cube topology is found to have favorable delay characteristics in light and moderate network traffic scenarios, while under heavy loading a completely connected topology turns out to be the most convenient one.**

Secure VPN; traffic analysis; hypercube; network performance

I. INTRODUCTION

In the field of Computer Security the development of a defense technique always stems from the analysis and understanding of the attack model which must be neutralized.

It is customary to classify security attacks into active and passive ones: the former category includes impersonation, forging/modifying content and denial of service, while the most studied form of passive attack is the interception and analysis of content. The latter can be counteracted by means of well known techniques, based on cryptography: the effectiveness of these solutions is very high.

Another kind of passive attack is traffic analysis. A lot of information can be gathered by an attacker which observes *timing*, *source*, *destination* and *size* of messages, even without any knowledge of their content. Neutralizing this kind of attack requires complex and expensive countermeasures, which are employed only in scenarios requiring the highest degrees of security.

This paper describes a part of the BlankNet project, being currently carried out at the University of Milano Bicocca. BlankNet defines the architecture of a secure VPN, which is capable of protecting its users against most kinds of attacks to traffic secrecy: in particular not only it does protect the content of messages by means of cryptographic techniques, but it is also designed to defeat traffic analysis attacks. The BlankNet VPN can be built upon any kind of network, ranging from an enterprise LAN to the Internet.

Much of the inspiration for the BlankNet architecture has come from our analysis of the characteristics of the Tor project [1]. This work is aimed at developing highly performing solutions solving the same kind of problems.

The subject of VPN topologies has been analyzed thoroughly in a context rather different from the present one, namely in the field of peer-to-peer networking: Chord is a very interesting example of these applications [2].

There is a huge literature on interconnection network for multiprocessors, where the binary cube and many other topologies have been studied [3][4]; this paper is indebted also to the research carried out on high speed space-division switching [5], particularly for what concerns some aspects of the performance analysis in Section V.

The rest of the paper is organized as follows: Section II contains a basic description of the BlankNet architecture; Section III sets up the key assumptions which allow a meaningful evaluation to be made; Section IV identifies a topology which is better than the ring and the complete connection in terms of packet delay in a lightly loaded network; Section V develops an analytical model which allows to extend the comparison of topologies to high loading situations; Section VI concludes the paper mentioning the issues to be faced when applying the results obtained to a concrete implementation scenario.

II. THE BLANKNET ARCHITECTURE

The security objectives of BlankNet can be obtained in a rather straightforward fashion, as described below.

All the stations transmit fixed size *packets* at fixed time intervals, towards destinations which are defined according to some predetermined rule. The payload of each packet may be a random bit sequence or an encrypted *message*, directed toward a final destination which may be different from the node which receives the packet.

Each node decodes the packets and processes the message headers, while the message payloads are protected by a second layer of coding, which ensures end-to-end secrecy; if a node is not the final destination of a message, encodes it again with a different key and forwards it according to a routing algorithm.

The encode/decode operations performed at each hop prevent an external observer from correlating the messages sent and received by a node and tracking the message paths: Figure 1 depicts the range of encoding operations.

The node architecture is represented in Figure 2, where a detail needs to be pointed out: while the routing buffer RQ is managed as a normal FIFO queue, the transmission buffer TQ must be seen as a set of FIFO queues, one for each possible packet destination.

 ISBN:978-1-61208-113-7

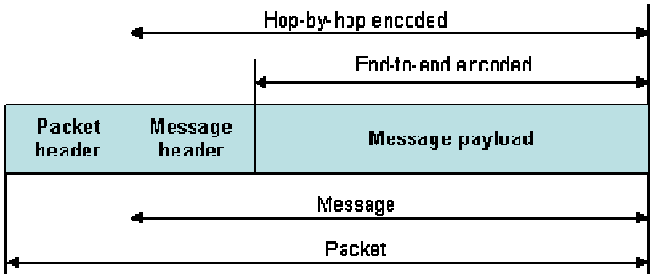


Figure 1. BlankNet message encoding

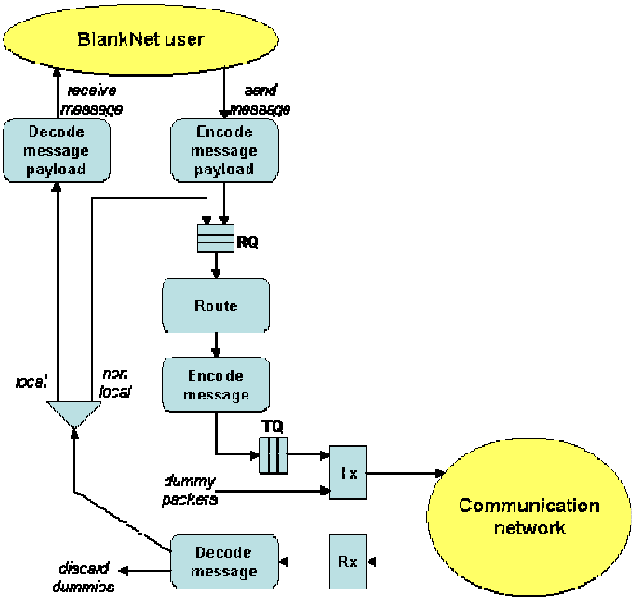


Figure 2. BlankNet node architecture

A key aspect of the BlankNet architecture is the definition of the allowed packet communication patterns: as it has been said before, the packets (which carry messages or random payloads) are sent to a predetermined set of nodes, at predetermined transmission times: we call this pattern a *virtual topology*. It is possible to implement a BlankNet using different virtual topologies, having different characteristics and performance.

In order to fix ideas, let us consider the unidirectional ring, depicted in Figure 3a, which is probably the simplest solution to implement. In this case, every node always transmits its packets to the same destination (node 1 to node 2, node 2 to node 3 etc). A message from node 1 to node 4 will be relayed by nodes 2 and 3 before reaching its destination: the traversal of intermediate hops is a penalty to be paid for obscuring the traffic patterns to external observers.

A different virtual topology is the complete connection depicted in Figure 3b. Here every node transmits its packets directly to the other nodes, but the transmissions must take turns (for instance node 1 sends its packets to the other nodes following the cyclic order: 2,3,4,5,2,3,4,5,2,…): in this case the performance penalty required for traffic pattern obfuscation is the time that messages wait for the transmission slot to the proper destination.

A bit of thinking suggests that these two topologies may have similar performance: in a network with N nodes with a ring topology, a message must cross N/2 (virtual) links in the average before reaching its destination, but it does not suffer from contention for the transmission time slot, while with a completely connected topology the message must wait, in the average, N/2 transmission slots before being sent (directly) to its destination.

It is however intuitive that, if we consider the delay experienced in a loaded network, with queues building up at each node, the complete connection should have a significant advantage with respect to the ring, because it does not employ network resources (transmission slots) for the intermediate hops.

These observations lead to two general questions, which are analyzed in this paper:

A) Can we find virtual topologies which are more efficient than the ring or the complete connection?

B) How do we model the performance of the various topologies in order to make a meaningful comparison in various operating conditions?

An answer to question A is given in Section IV, while Section V develops an analytical performance model.

III. BASIC SYSTEM PARAMETERS AND OPERATING MODES

As a first step towards the objective evaluation of the various virtual topologies, it is necessary to define precisely the application scenario where the comparisons are made. This definition involves the determination of some key parameters and operating modes of the VPN.

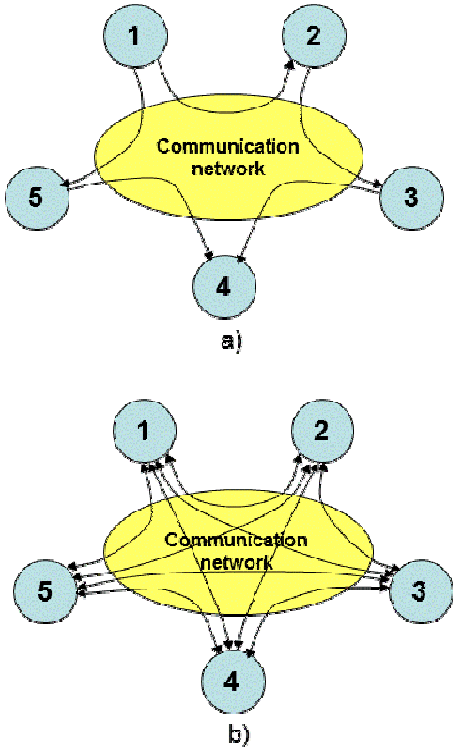


Figure 3. (a) unidirectional ring and (b) completely-connected virtual topologies

A. Transmission delays

In the first section of the paper we compared implicitly the time taken to cross N/2 virtual links to the time spent in waiting for the N/2-th transmission slot: however these two times do not need to be comparable. We shall use the following definitions:

- T_{slot} is the time between two successive packet transmissions performed by the Tx unit in Figure 2;
- T_{link} is the time needed to cross a virtual link, i.e. the time between the beginning of the transmission of a packet by Tx and the completion of its reception by the Rx block at the next node.

Note that, except for very particular implementations, only a fraction of the capacity of the access link towards the Communication Network is dedicated to the BlankNet VPN, while other applications can share the link: therefore, typically, T_{slot} is much larger than the transmission time of a packet. We shall assume that the transmission of BlankNet packets on the link has precedence over any other application, that is Tx preempts any other transmission process on the physical network interface.

T_{link} varies considerably depending on the underlying Communication Network: it can easily span two/three orders of magnitude going from a LAN to the Internet.

In order to characterize the various possible environments we define the parameter $r=T_{link} / T_{slot}$. In the following we shall consider only two cases, which in our view are the most meaningful:

- r close to 1, which may occur in an enterprise LAN environment: for instance, if we dedicate 10% of a 100 Mbps Ethernet link to the BlankNet, and transmit 10 kbit packets, T_{slot} would be 1 ms, which is comparable to T_{link} for an enterprise LAN;
- r around 100, which is a likely value in a public network environment: for instance, T_{link} may be 100 ms, and if we dedicate 10% of a 10 Mbps ADSL access link to the BlankNet, still transmitting 10 kbit packets, T_{slot} would be 1 ms.

Therefore our evaluations are carried out for r=1, r=10 and r=100.

B. Timing and synchronization

We do not make specific assumptions about the time taken by the various functional blocks in Figure 2, except that they are "fast enough". More precisely, for the purpose of our analysis, we suppose that, once a message is received, it can be transmitted in the next time slot (provided that no contention occurs at the TQ buffer). In other terms, T_{slot} is long enough to account for the local processing at each node.

Moreover, we make the simplifying hypotheses that r is an integer number, and that T_{slot} is the same for all the BlankNet nodes. Although these are clearly not very realistic, they do not have an impact on the relative performance of the various virtual topologies.

A further issue regards the relative synchronization of the various nodes in a BlankNet. The relevance of this problem can be shown by considering the complete connection virtual topology of Figure 3b. If all the nodes are completely independent, it is possible that all the other nodes transmit a packet to node 1 in the same time slot: this causes congestion at the input of node 1, adding a further delay to the message reception. This delay can be a large or negligible factor, with respect to the BlankNet performance, depending on the ratio of the VPN speed to the physical network speed.

If there is a global synchronization of transmission times similar problems can be effectively solved. However, for most part of following analysis, we shall assume that no global synchronization exists, but the congestion problem is negligible due to the high speed of the communication network and the limited degree of the majority of virtual topologies to be considered. In specific situations, we shall underline the advantages brought by synchronization.

C. Routing

The unidirectional ring is the only topology where there are no alternative routes to a destination; for instance, also in a completely connected topology it is possible to send a message to a different node, and then transmit it from there to the final node: this option could even be advantageous if the packet has a long time to wait for the scheduled transmission time through the direct link.

In the following we shall limit the routing options by means of these two rules:

1. a message is always transmitted to a node that is closer to destination in terms of (virtual) links to be crossed;
2. in case more than one node can be selected according to rule 1, the choice is made at random.

Rule 1 simplifies the routing operation and the analysis, and makes routing load-independent; rule 2 distributes traffic load evenly among all the possible routes.

IV. AN ALTERNATIVE VIRTUAL TOPOLOGY

In this section we will compare the different topologies in terms of the average time required to a message to reach its destination, if it does not find any other message ahead in the TQ buffers (light-loading). We define this metric as $D(0)=D_s+D_l$, where D_s is the time spent in the nodes waiting for the proper transmission slot, and D_l is the total time taken to cross all the links to destination. In the following we shall take T_{slot} as the time unit, therefore we have $D(0)=s+rd$, where s is the average total number of transmission slots spent waiting for transmission and d is the average number of links crossed. Note that, using terminology of graph theory, d is the average distance between two nodes of the topology graph.

It is easy to find out that for the ring we have $D(0)=rN/2$ and for the complete connection we have $D(0)=(N-2)/2+r$; because the complete connection is equal to the ring for $r=1$ and better in all the other cases, we shall not consider the ring in the following.

A. The binary cube

The binary cube topology comprises $N=2^k$ nodes, each one connected to k other nodes by bidirectional links. More precisely, every node can be identified by a bit string such as $b_{k-1}\ldots b_1 b_0$ and is directly connected to the k nodes $\underline{b}_{k-1}\ldots b_1 b_0$

 ISBN:978-1-61208-113-7

,…, $b_{k-1}…\underline{b_1}b_0$, $b_{k-1}…b_1\underline{b_0}$. Figure 4 depicts a binary cube with *N=16* (*k=4*).

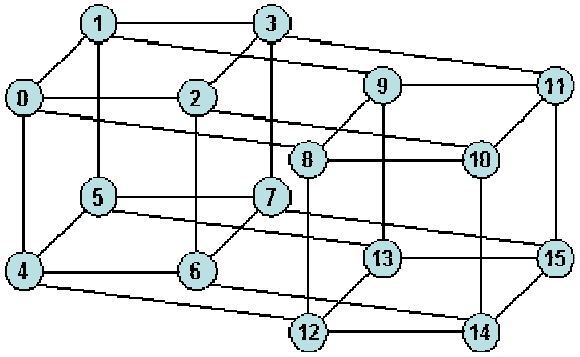


Figure 4. A four-dimensional binary cube

Starting from any node in a cube, we reach *k* new nodes crossing one link and $\binom{k}{i}$ new nodes after crossing the *i*-th link. Based on this consideration, it is simple to calculate the average distance in a binary cube, which is $d = \frac{k}{2}\frac{N}{N-1}$; moreover, because the nodes are supposed to work asynchronously, and we use rule 2 of the preceding section to decide on the next hop (instead of, for instance, sending the packet at the first time slot compatible with rule 1), we have $s = \frac{k-1}{2}d$. Therefore, for the binary cube, we obtain:

$$D(0) = \left(\frac{k-1}{2} + r\right)\frac{k}{2}\frac{N}{N-1} \qquad (1)$$

As we shall see in the next paragraph, a smarter routing rule and the synchronization of the nodes can create a significant performance advantage.

Table I reports the values of *D(0)* for the binary cube and the complete connection (CC) for various VPN sizes and *r* values.

TABLE I.

N	*r=1*		*r=10*		*r=100*	
	CC	*cube*	*CC*	*cube*	*CC*	*cube*
128	64	14,1	73	45,9	163	363,3
256	128	18,1	137	54,2	227	415,6
512	256	22,5	265	63,1	355	468,9
1024	512	27,5	521	72,6	611	523,0
2048	1024	33,0	1033	82,5	1123	577,8
4096	2048	39,0	2057	93,0	2147	633,2
8192	4096	45,5	4105	104,0	4195	689,1

It can be easily seen that the cube outperforms the complete connection in most use cases, the exceptions being for the combination of large *r* and small number of nodes (in grey on the table).

B. Synchronized operation of the binary cube

We can achieve a significant improvement of the cube performance if we synchronize globally the transmissions on the entire BlankNet. Let us assume that, in a specific time slot, all the transmissions are along virtual links corresponding to the same dimension of the cube, that is to say packets are sent towards destination nodes that differ from source nodes in the same bit position: Figure 5 depicts the four transmission phases (each lasting one T_{slot}) in the four-dimensional cube of Figure 4.

In order to understand the advantage of this global synchronization, let us consider at first the case of *r=1*. If *r=1*, it is guaranteed that a message, transmitted along a cube dimension, arrives at the beginning of the time slot corresponding to the next dimension in the transmission order. Therefore it is either transmitted immediately or waits for a time slot corresponding to a dimension it would not have crossed anyway.

As an example, assume the transmission order of Figure 5, and consider the path of a message entering at node 0 at the beginning of phase 2 and destined to node 11: transmission to node 2, wait one phase at node 2, transmission to node 10, transmission to node 11. This turns out to be the worst case, and requires *D(0)=k* phases, which compares quite favourably to the average values for the asynchronous cube.

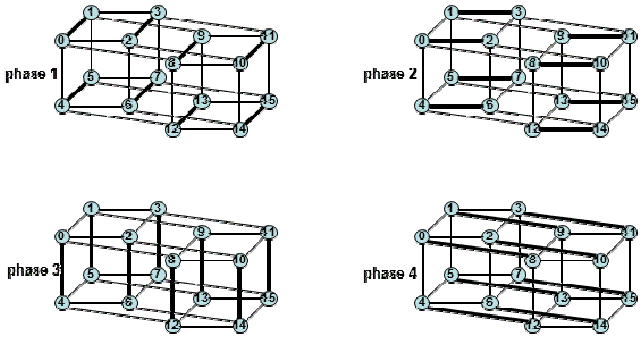


Figure 5. The four phases of synchronized transmission in a four-dimensional binary cube

Along these lines we can also derive an approximation for *D(0)* if *r>1*: it turns out that the same favourable synchronization occurs if *k* and *r* are mutually prime. In this case, a message from node *i* to node *j* would wait at most $s=k-d_{ij}$ slots (d_{ij} being the number of links to be crossed from *i* to *j*), therefore the delay formula for the synchronous binary cube can be upper bounded:

$$D(0) < k + (r-1)\frac{k}{2}\frac{N}{N-1} \qquad (2)$$

Of course the condition that *k* and *r* are mutually prime requires that r is deterministic, that is to say that the jitter introduced by the communication network is negligible with respect to T_{slot}. Because of this we assume that the synchronous behaviour is attainable only for small values of *r*. Table II compares the performance of the asynchronous cube and the above bound for the synchronous cube when

 ISBN:978-1-61208-113-7

r=1, 3 and *5* (we excluded the cases where *r* and *k* are not mutually prime).

TABLE II.

N	*r=1*		*r=3*		*r=5*	
	asyn	*syn*	*asyn*	*syn*	*asyn*	*syn*
128	14,1	7	21,2	14,1	28,2	21,1
256	18,1	8	26,1	16,0	34,1	24,1
512	22,5	9	31,6	---	40,6	27,0
1024	27,5	10	37,5	20,0	47,5	---
2048	33,0	11	44,0	22,0	55,0	33,0
4096	39,0	12	51,0	---	63,0	36,0
8192	45,5	13	58,5	26,0	71,5	39,0

V. AN ANALYTICAL MODEL

In order to take in account the effect of traffic, we use a simple queuing model of the node.

The hypotheses made on the operations of the BlankNet node lead us to analyze the behavior of the TQ buffer, which can be modeled as a set of *k* independent queues, each one corresponding to a transmission phase (or, equivalently, to a virtual link).

Because the purpose of our evaluations is the comparison of virtual topology, we can assume a fairly simple traffic model: a message is introduced in the network by each BlankNet user with probability λ at each T_{slot} (it does not make sense to suppose a faster transmission rate by the user because the maximum transmission speed at each node is one message per T_{slot}). Messages are directed with equal probability to all possible destinations in the VPN: as a consequence, because of the isotropy of the topologies that we are analyzing, all the virtual links are equally loaded.

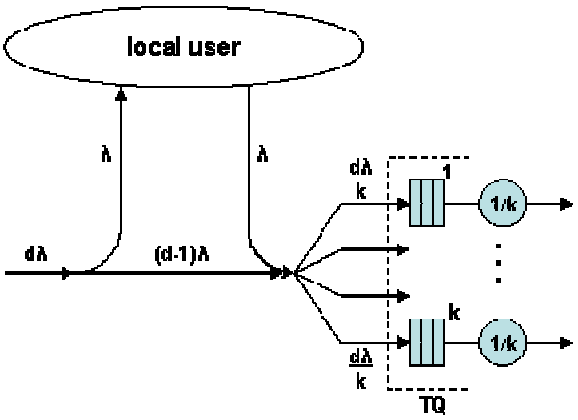


Figure 6. Distribution of the load on the virtual links in case of uniform traffic destination and isotropic topology

The load on each virtual link can be determined by a simple reasoning: every message introduced in the network crosses, in the average, *d* links before reaching its destination; therefore every physical link carries *dλ* messages per time unit (T_{slot}) in the average; because this load is equally divided into *k* virtual links, each having *1/k*-th of the total capacity, the load on every virtual link is *dλ* as well. Figure 6 depicts the assumptions above.

In the case of the binary cube, we approximate the behaviour of each virtual queue with a simple M/D/1 model, which gives the following expression for *W*, the average waiting time in each queue:

$$W = \frac{kd\lambda}{2(1-d\lambda)} \tag{2}$$

It is immediately seen that the VPN reaches saturation when $\lambda=1/d$. At this point we can add a load-dependent factor to the overall time required to transfer a message to its destination, obtaining $D(\lambda)=D_q(\lambda)+D_s+D_t$, where $D_q(\lambda)=dW$. Therefore, for the (asynchronous) binary cube $D(\lambda)$ has the following expression:

$$D(\lambda) = \left(\frac{\lambda k^2}{2 \cdot \left(2\frac{N-1}{N} - \lambda k \right)} + \frac{k-1}{2} + r \right) \frac{k}{2} \frac{N}{N-1} \tag{4}$$

For the complete connection we do not have transit traffic, and the queuing model that we use for each virtual queue is a more precise binomial/D/1: the queue has a service cycle lasting *N-1* time units, and an input process characterized by the following probability p_i of *i* arrivals (messages sent by the user) in a service cycle:

$$p_i = \binom{N-1}{i} \left[\frac{\lambda}{N-1} \right]^i \left[1 - \frac{\lambda}{N-1} \right]^{N-1-i} \tag{5}$$

The waiting time for this model is known to be [6]:

$$W = \frac{N-2}{N-1} \cdot \frac{(N-1)\lambda}{2 \cdot (1-\lambda)} \tag{6}$$

Hence the $D(\lambda)$ formula is:

$$D(\lambda) = \frac{N-2}{2} \left(1 + \frac{\lambda}{(1-\lambda)} \right) + r \tag{7}$$

Figure 7 and Figure 8 compare the $D(\lambda)$ functions for the cube and the complete connection for a small (*N=256*) and a large (*N=4096*) VPN, with *r=1*.

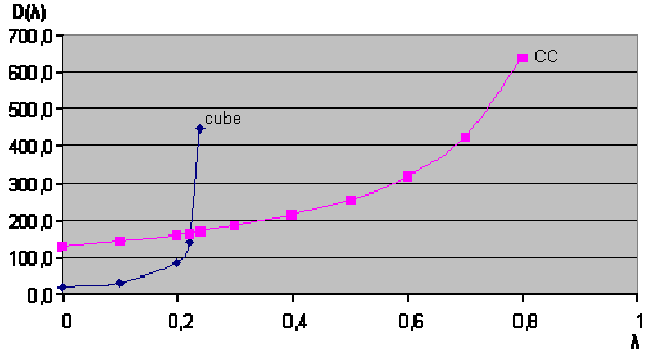


Figure 7. Delay function for *N=256, r=1*

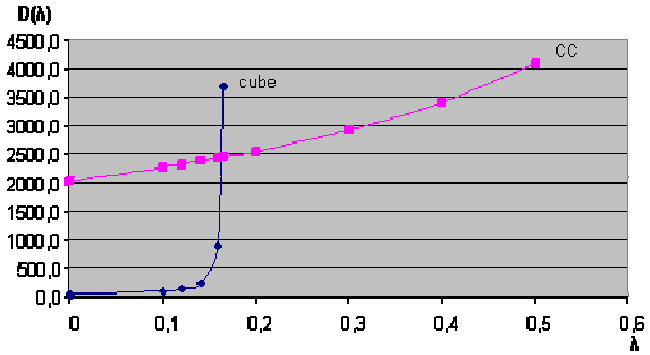


Figure 8. Delay function for *N=4096, r=1*

Figure 9 compares the two solutions with *r=100* for a large network (we know that for a small network the complete connection is superior even in light loading).

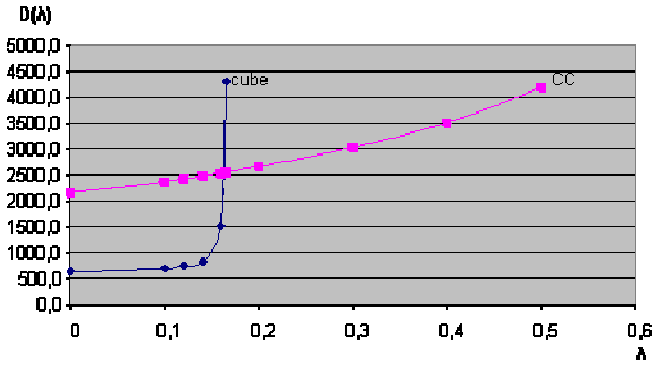


Figure 9. Delay function for *N=4096, r=100*

It is easy to see that, in all cases, the relative advantage of the two topologies changes drastically with the VPN loading, due to the effect of internal congestion which affects all topologies with *d>1*.

VI. Conclusion and Future Work

Our analysis has shown the existence of virtual topologies which perform much better than the ring and the complete connection as the basis of a BlankNet VPN.

However this advantage is present only for light-to-moderate loading conditions, after which we have to turn back to topologies with a very small average distance and to the complete connection in particular.

This work does not intend to suggest the use of the binary cube in practical situations, because it admits only configurations where *N* is a power of 2, a very unusual situation in a realistic application. However the characteristics of the binary cube suggest that several other candidate topologies exist, and in particular that we should work with graphs whose average distance grows logarithmically with the number of nodes (such as, for instance, the shuffle-exchange and its derivatives): these topologies, which can grow smoothly, shall be the objective of our further research.

The limitations of all these solutions under high loading suggest that we either implement a congestion control mechanism which prevents the BlankNet from working too close to the critical threshold $\lambda=1/d$, or that we develop a reconfiguration mechanism which allows the dynamic change of the virtual topology; while we are working on the first option, from a preliminary analysis the second one seems to introduce huge complexities without obtaining a significant payoff.

References

[1] www.torproject.org, checked on 06.01.2011

[2] Dabek, T. et al.: "Building Peer-to-Peer Systems with Chord, a Distributed Lookup Service," *Proc. 8th Workshop on Hot Topics in Operating Systems*, Elmau, Germany, May 2001.

[3] Bhuyan, L.N.: Qing Yang; Agrawal, D.P.; "Performance of multiprocessor interconnection networks," *Computer* , vol.22, no.2, pp.25-37, Feb 1989.

[4] Hayes, J.P.; Mudge, T.: "Hypercube supercomputers," *Proceedings of the IEEE* , vol.77, no.12, pp.1829-1841, Dec 1989.

[5] Karol, M.; Hluchyj, M.; Morgan, S.: "Input Versus Output Queueing on a Space-Division Packet Switch," *Communications, IEEE Transactions on* , vol.35, no.12, pp. 1347- 1356, Dec 1987.

[6] Meisling, T.: "Discrete-Time Queuing Theory," *Operations Research*, Vol. 6, no. 1, pp. 96-105, Jan. - Feb., 1958.

 ISBN:978-1-61208-113-7

Layer Optimization for DHT-based Peer-to-Peer Network

*Jun Li *, Cuilian Li, Zhaoxi Fang*
Department of Telecommunication
Zhejiang Wanli University
Ningbo, China
xxllj, licl@zwu.edu.cn, zhaoxifang@gmail.com

Haoyun Wang
College of Information Science & Technology
Nanjing Agricultural University
Nanjing, China
Wanghy@njau.edu.cn

***Abstract*—Hierarchical architecture has been found to facilitate effective search in P2P network and ensure system scalability in P2P application deployment. However, the lack of appropriate size ratio of super nodes layer and ordinary nodes layer makes the system search performance far from being optimal. Taking advantage of node heterogeneity, this paper presents a search delay model to characterize the two-layer P2P architecture using DHT at the top level. With the proposed models, optimal layer division can be achieved. Simulation analysis and numerical results validate the proposed model and solution.**

***Keywords*—Peer-to-Peer; Layer Division; Distribiuted Hash Table (DHT)**

I. INTRODUCTION

With the dramatic increase of Internet bandwidth and application of Peer-to-Peer (P2P) technology in Internet telephony, VoIP (Voice over IP) technology has developed rapidly and become a very popular communication vehicle due to its low cost and convenience to Internet users. Skype[1], the perfect combination of P2P and VoIP, sets a good paradigm to inspire a generation of P2P based solutions for satisfactory real-time multimedia services over Internet. Unlike file-sharing system with multiple replications of resources, VoIP user is unique in the system, thus it is more difficult and costly to locate the exact user, whereas the fast reach to the desired user is one of the key issues to user-perceived Quality of Service (QoS). In addition, VoIP P2P infrastructure is expected to accommodate millions of users and should be well scaled owing to its possible global application. Therefore, such a P2P infrastructure has to secure the specific requirements such as low lookup delay, high hit rate, and light workload.

Current decentralized P2P networks can be generally classified into two broad categories, structured and unstructured [2]. Purely unstructured P2P systems, such as earlier Gnutella [3], tend to either cost significant overheads or generate enormous query traffic in the exhaustive search, though they are characterized as high robustness and easy maintenance, which is the key reason that they are favored in the P2P paradigm. In contrast, structured P2P networks use distributed hash table (DHT) [4,5,6,7] for accurate object placement and lookup, but they are very sensitive to the dynamics of the network [8], owing to the fact that routing efficiency in DHTs is based on the consistent maintenance of routing tables. Typically, high dynamics bring dramatically high costs. Therefore, both of systems could not scale well, neither could achieve low lookup delay and high successful hit rate. Hierarchical structure that uses DHT to organize P2P network (say Chord) in the top level may address the performance problems such as scalability and resilience, motivated by the fact that participating nodes in P2P system differ a lot in uptime, bandwidth, etc. [9].

To further exploit the hierarchical DHT and develop the best performance for VoIP application, it is critical to address the problem of effective layer division. Thus, we propose the analysis model and give out the optimal size ratio between the number of super nodes (SNs) and ordinary nodes (ONs), taking into account the metric of total search delay under the constraints of SN capacity. Our study is expected to significantly reduce the mean lookup delay and effectively facilitate large-scale deployment of DHT based P2P lookup service by providing administrative autonomy nodes.

The rest of the paper is structured as follows. Section II discusses the related work. We present the models of total search delay in Section III, and solve the optimization problem in Section IV with simulation support. Finally, Section V concludes the paper.

II. RELATED WORK

There have been some studies on the hierarchical P2P network. Garces-Erice et al. [10] explored a general framework for hierarchical DHTs, instantiating Chord at the top level. They also analyzed and quantified the improvement in lookup performance of hierarchical Chord, considering the node failure. Yuh-Jzer Joung et al. [11] presented a two-layer structure Chord^2 to reduce maintenance costs, taking heterogeneity into account. Their work is distinguished to other hybrid architecture in the aspect that each layer forms an individual Chord ring.

On the basis of analyzing the promising super-node based P2P network, Yung-Ming Li et al. [12] firstly investigated the issue of sizing and grouping decisions from the perspective of P2P network organizers, due to the important role they play in determining network performance. Their work mainly focused on network scale determination, and grouping decisions were mentioned as well in the context of symmetric interconnection structures: isolated, chained and complete. However, the work doesn't aim for deployment. Since P2P network is self-organized with total autonomy, it is not very practical to predetermine or design the scale of the network. Motivated by the work of Li Xiao et al. [13], we focus on studying the grouping decision to improve performance of the P2P network and thus make it scalable, taking advantage of the node heterogeneity. Our work differs from theirs on two aspects: (1)

 ISBN:978-1-61208-113-7

We study DHT based upper layer, while they focused on the unstructured upper overlay instead. (2) We optimize the scheme by considering lookup delay with the capability constraint of the super node, while they tried to make a tradeoff between workload of super node and overall P2P network. Zoels et al. [14] proposed an analytical framework to analyze the same hierarchical P2P architecture as we are studying on. They evaluated the costs of the whole network as well as each participant, in order to determine an optimal layer division of a given system, similar to the way that work [13] applied. Recently, the authors of [15,16] further presented an analytical model for DHT-based two-tier P2P overlay and determined optimal fraction of superpeers in the system, aiming at minimizing the total traffic without overloading any peer. Complementary to their work, our goal is to minimize lookup delay of the system.

III. MODELING LOOKUP DELAY

A. Hierarchical structure model

In the considered two-layer hierarchical architecture, the participants are categorized as Ordinary Nodes (ONs) and Super Nodes (SNs). Those with longer uptime, better process capacity, network bandwidth and storage are elected as SN from the ONs. Each SN is responsible for its cluster and connects with all the nodes in the cluster. Meanwhile, the SN participates to form DHT-based SN network to facilitate upper layer searching (refer to Figure 1).

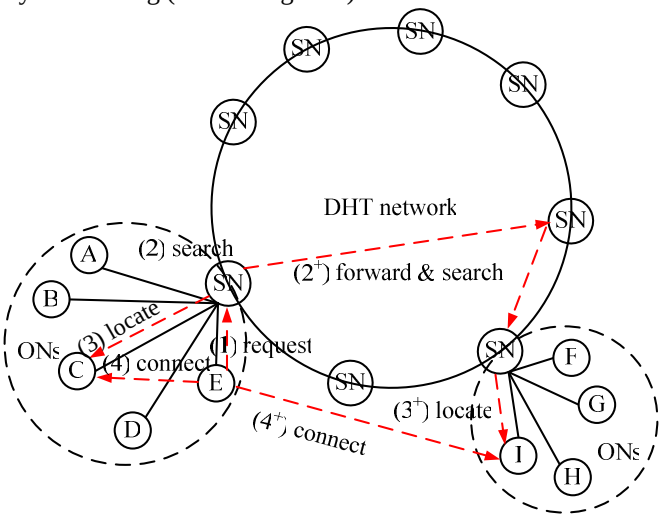


Figure 1. Search model in Hierarchical P2P network

Only the SN maintains up-to-date information on all resources (nodes in the context of VoIP) available in the cluster. Every search query is generated at one of the nodes (including SN) in the cluster, and first processed at the local SN on a first-come, first-served basis. For a specific query, SN first examines the resource in its cluster. If not satisfied, the query will be forwarded to other SNs and searched in the upper SN network. Figure 1 depicts the search operations of a hierarchical P2P network. Node E sends a query to its responsible SN. If the target node is within the cluster, say node C, SN will locate it by examining its resource list. If the query could not be satisfied in the cluster, say target node I, SN will search in the upper DHT-based P2P network (instantiating Chord in this paper). Once the desired node is located, direct connection could be established between the two nodes.

Consider a P2P hierarchical network with N participating nodes, in which N_{SN} nodes are SNs and N_{ON} nodes are ONs. Let η denote the layer size ratio, that is $\eta = N_{ON} / N_{SN}$, the number of nodes that each SN takes care of. We have $\frac{N_{SN}}{N} = \frac{1}{1+\eta}$, $\frac{N_{ON}}{N} = \frac{\eta}{1+\eta}$, which denote the ratio of the number of SNs and ONs to that of the participating nodes, respectively. Without loss of generality, we assume that all the nodes in the same category (SN or ON) are statistically identical.

In this section, we characterize the impact of layer size ratio η on probability of SN failure and then model total search delay under the constraint of available SN capacity.

B. Probability of SN failure

In [6], node availability is calculated by dividing the number of probes that a host responds to by the total number of probes in Overnet, which is structured on a DHT called Kademlia. According to their measurement results, assuming x is node availability, the CDF (Cumulative Distribution Function) of node availability could be well modeled by power distribution of the form x^k, where k mostly falls in the interval of (0.3,1) within normal observation period (illustrated in Figure 2).

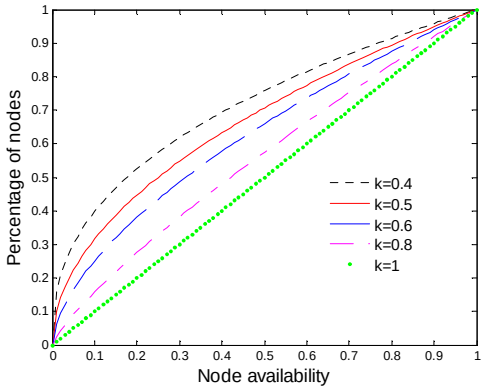


Figure 2. Node availability modeled by power distribution with varying observation time period

That is, the percentage of nodes whose node availability not exceeding value x is x^k. Equally, the percentage of nodes whose node failure exceeds value (1-x) is x^k. Given layer size ratio η, the ratio of the number of SNs to that of all the participating nodes is $\frac{N_{SN}}{N} = \frac{1}{1+\eta}$. Since SNs typically have longer uptime than ONs and therefore achieve higher node availability, we have

$$1 - \frac{1}{1+\eta} = x^k, \text{ and } x = \left(\frac{\eta}{1+\eta}\right)^{\frac{1}{k}}$$

That is, given η, the minimum node availability for an SN is $(\eta/(1+\eta))^{\frac{1}{k}}$. Accordingly, failure rate of each SN will not exceed $1-(\eta/(1+\eta))^{\frac{1}{k}}$. In the worst case, all SNs have

failure rate $p = 1-(\eta/(1+\eta))^{\frac{1}{k}}$.

C. Modeling and analyzing search delay

In SN based hierarchical structure, search delay consists of two parts: (1) queuing delay at the SNs, (2) propagation delay during search operation, including propagation delay between ON and its responsible SN, and search propagation delay in SN overlay network. As locality-aware mechanism has been studied and widely applied in P2P paradigm [17], we can introduce locality-awareness upon ONs joining SN and the mean latency between an ON and its responsible SN $E[T_{Intra}]$ is trivial compared to the mean latency between SNs in the upper layer. As to search propagation delay, we define the metric of lookup hops in SN upper network as the hops taken from sending query to receiving response, instead of the mostly measured hops from sending query to reaching the responsible node in flat P2P network, for the sake of fair comparison. Thus, one more hop in P2P network to forward the result back would be included.

We characterize the mean lookup hops in the upper SN network in this subsection. Then we model queuing delay at an SN. Finally the total search delay is proposed.

1) Modeling lookup hops in upper SN network

Taking Chord for instantiation as the upper DHT based SN network, we quantify the mean lookup latency in Chord network according to Garces-Erice's work [10], and get the result by calculating in Matlab, as illustrated in Figure 3.

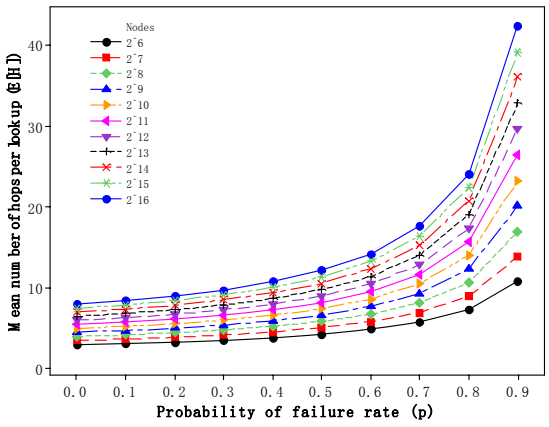


Figure 3. Mean number of hops per lookup

Let $E[hop]$ denote the mean overall lookup hops taken in the SN network, i.e. $E[hop] = E[H]+1$, where $E[H]$ is the mean lookup hops from the requesting SN to the target SN, and we add one hop for the target SN tracking back to the querying SN. $E[H]$ is obtained according to the work [10] with Chord population $N_{SN} = N/(1+\eta)$ in the upper layer.

Apparently, $E[H]$ varies with N, node failure rate p and layer size ratio η. Since failure rate p relies on η value, $E[H]$ is the function of η when N is given.

2) Modeling queuing delay at SN

Standard queuing model is applied to evaluate the delay occurring at each SN. The service time for search process at SN is assumed to follow an exponential distribution with service rate μ. Previous research [18] suggests that Poisson process is valid for modeling arrivals of user-initiated requests. Dang et al. [19] also provide evidence that VoIP call arrival forms Poisson process. We therefore assume that requests follow a Poisson process. Since requests from all the nodes are independent Poisson processes, the aggregate request arrival at an SN is also a Poisson process and the search at SN can be modeled as an M/M/1 queue.

Let f_q be the query arrival rate from each node, the aggregate requests arriving at each SN consist of those from its cluster $\lambda_i = f_q(\eta+1)$ and those forwarded traffic in the upper DHT network λ_e. The aggregate request arrival rate is $\lambda = \lambda_i + \lambda_e$. We assume that each SN shares the overall lookup workload in Chord evenly, since SNs are distributed uniformly and independently in the identifier space in light of the Chord algorithm. Thus,

$$\lambda_e = \frac{Nf_q(1-\eta/(N-1))E[hop]}{N_{SN}} = f_q(1+\eta)(1-\frac{\eta}{N-1})E[hop]$$

Where $\eta/(N-1)$ is the expected node availability within the local cluster, as the probability of the target node which the initiated node seeks for is uniformly distributed among all the participating nodes. Therefore, $1-\eta/(N-1)$ is the expected probability that the query has to be forwarded and circulated in the upper SN network. The expected sojourn time (queuing delay) at an SN is as follows

$$E[T_w] = \frac{1}{\mu-\lambda} = \frac{1}{\mu - f_q(1+\eta)(1+(1-\eta/(N-1))(E[H]+1))} \quad (1)$$

with the constraint of $\mu/\lambda > 1$.

Let $f(\eta) = (1+\eta)(1+(1-\eta/(N-1))(E[H]+1))$, then we have $f(\eta) < \mu/f_q$. We examine the derivation of $f(\eta)$, i.e.

$$f'(\eta) = 1+(1-\frac{1+2\eta}{N-1})(E[hop]) + E'[hop](1+\eta)(1-\frac{\eta}{N-1}).$$

As illustrated in Figure 3, $E[H]$ is increasing with respect to network scale. Furthermore, the increment of lookup hops with different failure rate (especially failure rate <0.5, which is easily achieved in SN upper network) is similar to that of failure free $p = 0$. We assume that

$$E'[hop] \doteq \left(\frac{1}{2}\log\frac{N}{1+\eta}+1\right)' = -\frac{1}{2\ln 2(1+\eta)}, \text{ thus we have}$$

$$f'(\eta) = 1+(1-\frac{1+2\eta}{N-1})E[hop] - \frac{1}{2\ln 2}(1-\frac{\eta}{N-1})$$

$$> (1-\frac{1+2\eta}{N-1})E[hop] + \frac{\eta}{N-1} > 0$$

That is, $f(\eta)$ is monotonously increasing with respect to η. For $\eta < \eta_{\max}$, where $\eta_{\max}$ is the upper bound of value η, $E[T_w] = \frac{1}{\mu - f_q \times f(\eta)}$ is thus monotonously increasing with respect to η as well. We can have the same result as illustrated in Figure 4, where $N = 2^{16}$, $k = 0.7$, $f_q = 1/60$, $\mu = 15, 20, 25, 30$ respectively, for instantiation.

Normally, available SN service capacity μ and request arrival rate of each node f_q are given, thus we can find upper

bound $\eta_{\max}$, according to $f(\eta) = \mu / f_q$. Let $\eta < \eta_{\max}$, then the constraint $f(\eta) < \mu / f_q$ is satisfied. Herein, $\eta_{\max}$ is a critical parameter, as queuing delay increases sharply when $\eta \to \eta_{\max}$.

Numerical Results: Some numerical results of $\eta_{\max}$ calculation are presented as follows. Set failure parameter k=0.7, N= 2048, 8192 and 65536 respectively. We vary the value of μ / f_q to determine the corresponding upper bound $\eta_{\max}$ and the results are illustrated in Figure 5. With the identical μ / f_q, the larger the network scale, the least the $\eta_{\max}$ can be obtained, leading to the narrower interval for choosing η. As the P2P network is toward expanding dramatically over time suggested by the present P2P networks, the issue should be seriously taken into account when designing P2P network architecture or dynamically adjusting η in runtime. In addition, the upper bound $\eta_{\max}$ is increasing with respect to the value μ / f_q, which indicates that better SN capacity can enlarge the possible η value interval if the request rate is fixed.

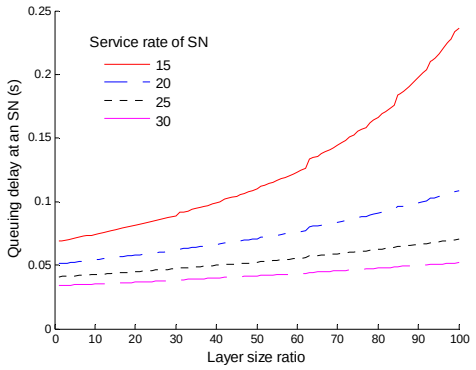


Figure 4. Queuing delay at SN vs. η with varying μ

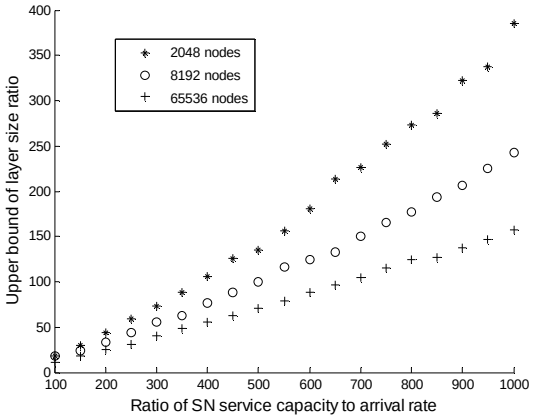


Figure 5. $\eta_{\max}$ vs. given μ / f_q

3) Total lookup delay

Assuming that the propagation between ON and SN within a cluster is negligible compared to the propagation between SNs, i.e. $E[T_{Intra}] \approx 0$, as ON joins SN with locality-awareness, the total mean lookup delay can be expressed as follows:

$$E[T] = \frac{\eta}{N-1} E[T_w] + (1 - \frac{\eta}{N-1})(E[T_w] + (E[T_w] + D)(E[H]+1)) \quad (2)$$

Where $\frac{\eta}{N-1} E[T_w]$ is for the query which is satisfied within the local cluster, thus total lookup delay is the queuing delay $E[T_w]$ at the responsible SN with the hit rate $\frac{\eta}{N-1}$. Otherwise, the query will be forwarded by the SN and searched in the upper SN network. In addition to the queuing delay at the responsible SN, each hop in the upper DHT search network will take $E[T_w] + D$, where D is the latency between two arbitrary nodes in the P2P network. Altogether there are $(E[H]+1)$ hops in average as aforementioned. We therefore have total lookup delay $(E[T_w] + (E[T_w] + D)(E[H]+1))$ for those searching in the external of the cluster.

Substitute $E[T_w]$ to Equation (2) and rewrite it. We have

$$E[T] = \frac{1 + (1 - \frac{\eta}{N-1})(E[H]+1)}{\mu - f_q(1+\eta)(1 + (1 - \frac{\eta}{N-1})(E[H]+1))} \quad (3)$$

$$+ D(1 - \frac{\eta}{N-1})(E[H]+1) = T_q + T_p$$

We define two types of delays: total queuing delay at SNs T_q and total search propagation delay T_p, which is incurred by the hopping time in the SN network.

IV. SOLUTION TO OPTIMAL LAYER SIZE RATIO

In this section, we present the optimization problem and solve the optimal layer size ratio according to the above analysis, demonstrated by numerical results in Matlab. Simulation is conducted to support the proposed model.

A. Problem statement

The problem of seeking for optimal layer size ratio for the hierarchical architecture is to find the optimal ratio of the number of ONs to the number of SNs, which could achieve the least total lookup time under the constraint of the available SN's capacity. That is, it is an optimization problem stated as follows:

$$\begin{cases} \min & E[T] \\ s.t. & \lambda < \mu \end{cases} \quad i.e. \quad \begin{cases} \min & E[T] \\ s.t. & \eta < \eta_{\max} \end{cases} \quad (4)$$

Herein, μ is the average available service capacity of SN. From the Equation (3), $E[T]$ is dependent of $E[H]$, μ, η, f_q and D. According to the previous analysis, $E[H]$ relies on layer size ratio η solely for a specific network with N nodes. Typically, parameters of μ, D and f_q are given for a specific system, so $E[T]$ is the function of η. Intuitively, we can find the optimal value of η at the minimum of $E[T]$.

B. Solving optimal layer size ratio

King[20] estimates RTT (Round Trip Time) between any two hosts in the Internet by estimating the RTT between their domain name servers. Since the edges in P2P network are not physical communication links, but instead only virtual links between the peers, the nodes could be geographically

 ISBN:978-1-61208-113-7

dispersed. Therefore, the latency between two P2P nodes is just as the latency between any arbitrary hosts, and we have the mean delay between two SNs in P2P network $D = 0.078$ second, by analyzing the current available data.

For specific network with given node capacity limitation, since parameters are predetermined by the system, we therefore have the optimal η value to achieve the minimum lookup delay under the constraint of SN capacity. We examine the mean overall lookup hops $E[T]$ with the varying layer size ratio η, which is the unique impacting factor if network size N is determined. Take a network with 2^{16} nodes for instance, and set typical parameter values as follows: $k = 0.7, f_q = 1/60, D = 0.078, \mu = 15$. Figure 6 illustrates the trends of the mean overall lookup delay with the varying η. Total search delay $E[T]$ decreases when η is low, while increases as η goes higher. Once η is approaching $\eta_{\max}$, $E[T]$ increases steeply as queuing delay at SNs is increasing dramatically.

In addition, we can find that SN search propagation delay $D \times E[hop]$ decreases with respect to η, and the curve of the value turns to be flatter as η increases. The reason lies in the fact that as η increases, the number of SN $N_{SN} = N/(1+\eta)$ decreases. Fewer SNs are involved in search process in the upper SN layer, leading to fewer search hops and higher search efficiency. Besides, larger η means fewer SNs, which results in better capacity and lower failure rate of SNs, further reducing lookup hops as illustrated in Figure 3. Therefore, $E[H]$ decreases with the increase of η, so does $E[hop]$.

However, the total queuing delay is dominated in the share of total lookup delay when η is large, especially in the case that available capacity is low compared with the arrival rate, the curve is more dependent on the queuing delay. While queuing delay is increasing with respect to η, and approaching infinite theoretically when $\eta \to \eta_{\max}$.

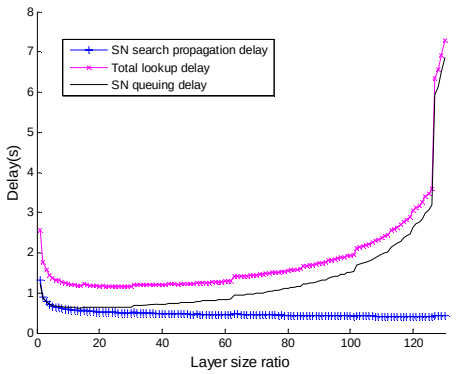


Figure 6. Delay varying with respect to η

We can find that the curve of aggregate search delay is concave and there exists a minimum value of delay with corresponding η. In this case, for example, we observe that promising η falls in the interval of [20, 50]. The optimal value is $\eta = 30$ and the least total search delay is only around 1.15 seconds accordingly, even SN is with poor capacity in this example.

It is noticeable that the curve is not smoothly decreasing as expected since a small increase can be seen when the number of SNs $N/(1+\eta)$ just exceeds a binary exponential value 2^i. The underlying reason is as follows: once the number of SNs crosses the next power of 2, the round function makes the curve incontinuous and the mean overall lookup hops slightly increase at the point.

Finally, we examined a variety of instances of parameters such as network size N, arrival rate and capacity of SN. The results keep similar, except that $\eta_{\max}$ is changing and leading to the varying scope and shift of the solution interval for the optimal η.

C. Simulation validation

In order to validate the established model, we modify and construct the two-layer P2P architecture using Chord as the top searching network based on P2PSim [21] simulation environment. In the P2P architecture, SNs are organized into Chord ring on the top layer, whereas the ONs directly connect to SNs with locality awareness. SN's available service capacity is not distinctly specified.

We tested the mean lookup delay in modified P2Psim environment with the maximum overall node population, i.e. N=2048 due to the limitation of the P2PSim. We set $f_q = 1/60, \mu = 10$. The simulation results of mean search delay mostly fall between SN search propagation delay and total search delay calculated by our model, as illustrated in Figure 7. Since SN capacity is not limited, the queuing delay could be very low. It is shown that simulation result can support our modeling and validate the solution.

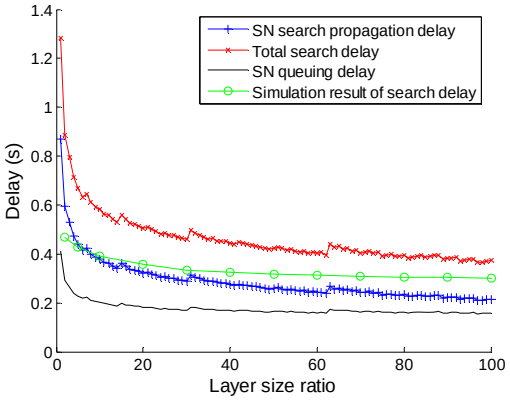


Figure 7. Simulation and modeling results of overall mean lookup hops vs. layer size ratio

V. CONCLUSION AND FUTURE WORK

Two-layer P2P organization with DHT (say Chord) at the top level is a promising hierarchy to improve performance of P2P network and satisfy numerous P2P applications, especially for VoIP deployment. This paper proposes the performance models by analyzing total lookup delay with queuing delay at SNs and propagation delay in upper DHT based searching network. Based on the models, an optimal

 ISBN:978-1-61208-113-7

layer size ratio can be achieved to minimize the overall lookup delay under the constraint of available SN service capacity by means of numerical results. Our simulation results support the proposed models and solutions. The crucial issues such as system maintenance and churn disposal will be studied as our future work.

ACKNOWLEDGMENT

The paper is partially supported by Zhejiang Natural Science Foundation under Grant Y1080935 and Y1101123. Ningbo Natural Science Foundation under Grant 2010A610121 and 2010A610174

REFERENCES

[1] S. Baset and H. Schulzrinne, "An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol", Proceedings of 25th IEEE International Conference on Computer Communications, INFOCOM 2006, Barcelona, Spain, pp. 1-11, 2006.

[2] EK. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A survey and comparison of peer-to-peer overlay network schemes," *Communications Surveys & Tutorials, IEEE*, vol. 7, no. 2, pp. 72-93, 2005.

[3] "Gnutella," http://gnutella.wego.com/, 2003. [retrieved: November 3, 2010]

[4] S. Ratnasamy, P. Francis, M. Handley et al., "A scalable content-addressable network," Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (*SIGCOMM'01*), San Diego, California, USA, pp. 161-172, 2001.

[5] I. Stoica, R. Morris, D. Liben-Nowell, D. Karger, M. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: a scalable peer-to-peer lookup service for Internet applications", ACM SIGCOMM 2001, 2001, pp. 149-160.

[6] P. Maymounkov, and D. Mazieres, "Kademlia: A peer-to-peer information system based on the XOR metric", Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS'02), pp. 53-65, 2002.

[7] B. Zhao, J. Kubiatowicz, and A. Joseph, "Tapestry: an infrastructure for fault-tolerant wide-area location and routing," University of California, Berkeley, Report No. UCB/CSD-01-1141, April 2001.

[8] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz, "Handling churn in a DHT", Proceedings of the 2004 USENIX Annual Technical Conference (USENIX'04), Boston, Massachusetts, USA, pp. 127-140, 2004.

[9] K. P. Gummadi, R. J. Dunn, S. Saroiu, and et al., "Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload", Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP-19), Bolton Landing, NY, pp. 314-329, 2003.

[10] L. Garces-Erice, E. W. Biersack, K. W. Ross, P. A. Felber, and G. Urvoy-Keller, "Hierarchical P2P systems", Proceedings of ACM/IFIP International Conference on Parallel and Distributed Computing (Euro-Par), 2003, pp. 1230-1239.

[11] Y. J. Joung, and J. C. Wang, "Chord2: A two-layer Chord for reducing maintenance overhead via heterogeneity," *Computer Networks*, vol. 51(2007), pp. 712-731, 2007.

[12] Y. M. Li, Y. Tan, and Y. P. Zhou, "Analysis of Scale Effects in Peer-to-Peer Networks," *Networking, IEEE/ACM Transactions on*, vol. 16, no. 3, pp. 590-602, 2008.

[13] X. Li, Z. Zhuang, and Y. H. Liu, "Dynamic layer management in superpeer architectures," *IEEE Transactions on Parallel and Distributed Systems*, vol. 16, no. 11, pp. 1078-1091, 2005.

[14] S. Zoels, Z. Despotovic, and W. Kellerer, "Cost-Based Analysis of Hierarchical DHT Design", Sixth IEEE International Conference on Peer-to-Peer Computing, 2006 (P2P 2006), 2006, pp. 233-239.

[15] S. Zoels, Z. Despotovic, and W. Kellerer, "On hierarchical DHT systems–An analytical approach for optimal designs," *Computer Communications*, vol. 31, no. 3, pp. 576-590, 2008.

[16] S. Zoels, Q. Hofstatter, Z. Despotovic, and W. Kellerer, "Achieving and maintaining cost-optimal operation of a hierarchical DHT system", IEEE International Conference on Communications, 2009.ICC'09. IEEE, 2009, pp. 1-6.

[17] B. Y. Zhao, A. Joseph, J. Kubiatowicz, "Locality aware mechanisms for large-scale networks", In Proceedings of the FuDiCo'02, pp. 80-83, 2002.

[18] V. Paxson, and S. Floyd, "Wide area traffic: the failure of Poisson modeling," IEEE/ACM Transactions on Networking (TON), vol. 3, No. 3, pp. 226-244, 1995.

[19] T. D. Dang, B. Sonkoly, and S. Molnar, "Fractal Analysis and Modeling of VoIP Traffic," NETWORKS, pp. 13-16, 2004.

[20] K. P. Gummadi, S. Saroiu, and S. D. Gfibble, "King: Estimating Latency between Arbitrary Internet End Hosts", Proceeding of SIGCOMM Internet Measurement Workshop, Marseille, France, 2002

[21] "p2psim," http://pdos.csail.mit.edu/p2psim/. [retrieved: November 16, 2010]

 ISBN:978-1-61208-113-7

Performance Evaluation of Split Connection Methods for Session-based Group-oriented Communications

Hiroshi Emina, Hiroyuki Koga
Dept. of Information and Media Engineering
University of Kitakyushu, Japan
Email: {emihiro, koga}@net.is.env.kitakyu-u.ac.jp

Masayoshi Shimamura, Takeshi Ikenaga
Network Design Research Center
Kyushu Institute of Technology, Japan
Email: {shimamura, ike}@ndrc.kyutech.ac.jp

***Abstract*—Group-oriented communication services have become more attractive due to the diversifying demands of Internet users. In these services, the bandwidth available to the link in the network is inefficiently consumed by multiple TCP connections and these individual connections compete against each other. Thus, current group-oriented services cannot use network resources efficiently. To achieve efficient group-oriented communication, we previously proposed a split connection method, which introduced session agents to eliminate redundant connections with a packet caching function. Furthermore, we showed the effectiveness of the proposed method and raised the cache utilization issue of the session agents under high traffic loads. In this paper, we propose packet transmission control methods to improve the cache efficiency of the session agents, especially under high traffic loads.**

***Keywords*-group-oriented communication, session management, connection control, packet transmission control**

I. Introduction

New communication services have become more attractive because of the diversifying demands of Internet users [1]. A one-to-one communication style, based on the client-server model, has been widely used for web and mail services. However, in recent years, a group-oriented communication style, based on a peer-to-peer model, has realized content exchange services, such as file sharing and online game services, among multiple users. In such group-oriented communication services, users need to share session information with other group members to manage the session. This session information consists of member IP addresses and group identifiers. In addition, content shared or exchanged by users in a group must be reliably delivered to other group members.

The current Internet has a significant difficulty in providing these services efficiently, which is in the reliable content sharing among group members. For example, consider that a sender establishes individual TCP connections with correspondent receivers in a group using current Internet technologies such as the Application Layer Multicast (ALM). In this case, the sender establishes additional TCP connections as the number of receivers increase. Consequently, the bandwidth available to the link is inefficiently consumed by multiple TCP connections that deliver the same data and these individual connections compete with each other. Thus, group-oriented communication services on the current Internet use network resources inefficiently.

To achieve efficient group-oriented communication services, we previously proposed a split connection method, which introduced a session agent with a packet caching function on intelligent intermediate nodes to eliminate the redundant connections [2]. Although this method improves the efficiency of network resource usage, it has an issue of cache utilization of the session agent under high traffic loads. Therefore, in this paper, we propose packet transmission control methods to improve the session agent cache efficiency, especially under high traffic loads. Furthermore, we show the efficiency of the proposed method by means of simulations.

II. Related work

Many researchers have actively studied technologies to provide group-oriented communications. In this section, we describe IP multicast and ALM mechanisms as typical group-oriented communication technologies. Then, we introduce the split connection method from our previous work.

A. Multicast

Applications based on IP multicast technology provide one-to-many communication services, such as real-time audio and video streaming services over the Internet. Although IP multicast provides flexible and efficient communication, it commonly supports unreliable datagram services. Therefore, IP multicast does not meet the need for reliable group-oriented communication service.

TCP-based ALM technology is commonly used to provide reliable group-oriented communications. In ALM networks, a sender transmits data to other members on the basis of a one-to-one communication style, namely, each member duplicates and forwards data packets to other members. To achieve effective communication services, researchers have proposed a number of ALM architectures [3]–[7]. Although ALM provides flexible group-oriented communication, it results in an inefficient use of network resources. To describe the problem of current group-oriented communication, we

 ISBN:978-1-61208-113-7

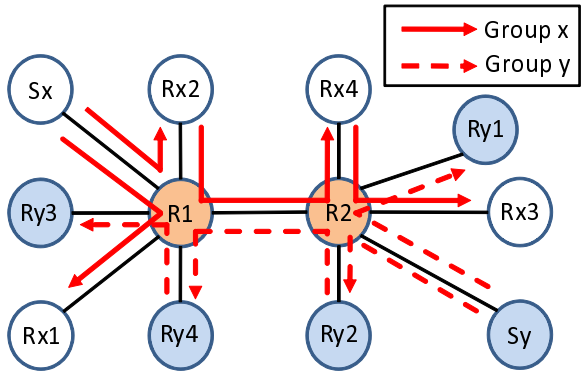


Figure 1. Group-oriented communications with TCP-based ALM

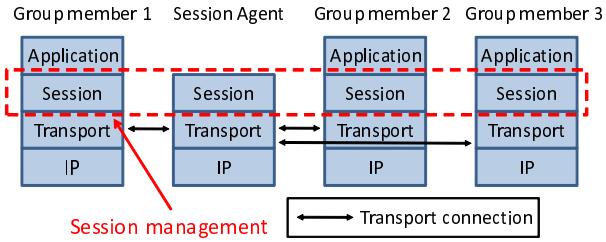


Figure 2. Layered structure of split connection method

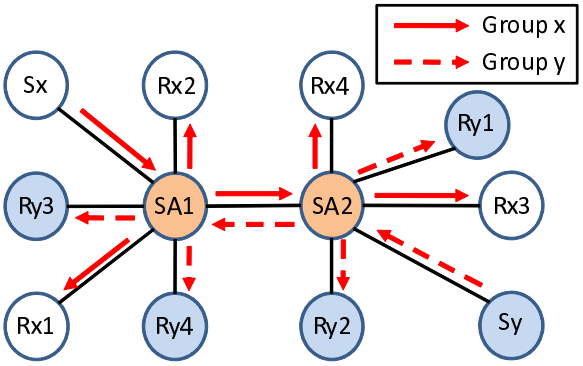


Figure 3. Group-oriented communications with a split connection method

show an example of group-oriented communications with TCP-based ALM in Figure 1. In this figure, two communication groups (x and y) exist, each consisting of one TCP sender and four TCP receivers. In group x, when sender Sx wants to transmit data to the four receivers, it needs to establish individual TCP connections with receivers Rx1 and Rx2. Rx2 then establishes TCP connection with Rx4, and Rx4 establishes another with Rx3 to forward the data. Similarly, in group y, each member also establishes multiple individual TCP connections with other members in the same group. In this situation, the bandwidth available to the link between Sx and R1, Sy and R2 are inefficiently consumed by multiple TCP connections that deliver the same data and these TCP connections compete with each other.

This inefficient group-oriented communication is caused by the management of the relationship between the session and transport connection in the current Internet architecture. To be more precise, the current Internet architecture does not have a session management layer and leaves session management to the application layer. Thus, each member must manage the session among the members on their application layer in group-oriented communications. Thus, each member establishes individual transport connections with other members on an end-to-end basis, and the number of connections increases proportionally to the number of receivers.

B. Split connection method

The current Internet architecture has a significant issue with achieving efficient group-oriented communications because of the lack of intelligent session management. To resolve this issue, in a previous work, we proposed a split connection method [2] based on a new network architecture concept [8]. In this method, we introduced an intelligent intermediate node, called a session agent. Figure 2 shows the layered structure of proposed architecture. In this figure, we assume that three group members and one session agent form a group. In this architecture, the session layer is inserted between the application and transport layers to efficiently manage a session among group members. The session layer provides several functions to manage join and leave requests from group members and to keep an end-to-end semantics among group members, which are provided by the application layer in the traditional architecture. This session management in the session layer enables this architecture to handle transport layer connections flexibly. More specifically, the session agent divides end-to-end TCP connections into multiple TCP connections partitioned by each session agent with packet caching, duplicating and forwarding functions. Consequently, the proposed method can improve network efficiency by aggregating redundant multiple TCP connections into one TCP connection in the same link.

Figure 3 depicts an example of group-oriented communications with the proposed method under conditions similar to those in Figure 1. In group x, all end-to-end TCP connections are split into multiple TCP connections. As a result, the sender Sx can aggregate two TCP connections into one connection and transmit a packet to session agent SA1 through the connection. After that, SA1 duplicates the packet into three flows, to the next session agent SA2, and group members Rx1 and Rx2. Similarly, SA2 duplicates and forwards the received packets to Rx3 and Rx4. For group y, the splitting of end-to-end connections can be performed similar to group x. As a consequence, the proposed method reduces redundant multiple TCP connections and improves network efficiency.

 ISBN:978-1-61208-113-7

III. PACKET TRANSMISSION CONTROL METHOD

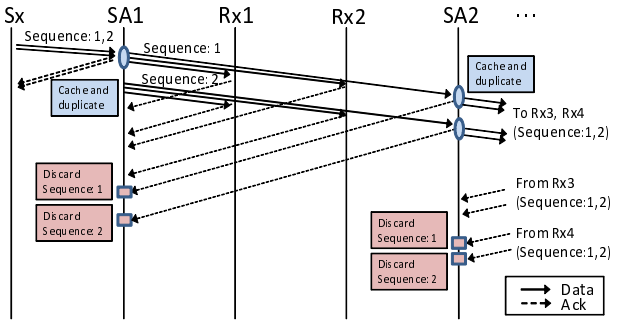
The split connection method improves network efficiency because of the aggregation of redundant TCP connections into one connection. However, this method has an issue in the packet cache utilization of session agents, especially for high traffic loads. Therefore, a sender should take into account the size of the packet cache of session agents when sending data packets. Therefore, to improve packet cache efficiency, we propose two packet transmission control methods: node-to-node and end-to-end.

A. Node-to-Node control

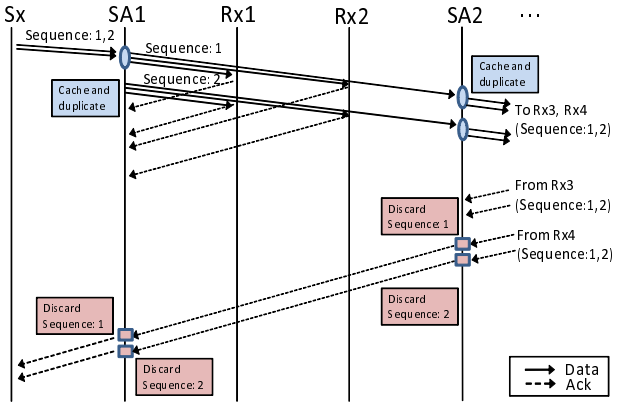
In node-to-node packet transmission control, each TCP connection partitioned by session agents controls its transmission rate independently, namely, a receiver (including session agents) requests the next data packet to a sender immediately after it receives data packets, regardless of the transmission state to other group member. Therefore, each sender (including group members and session agents) can optimize the available link bandwidth. Figure 4(a) illustrates the operation of session agents when a member in group x communicates with other members in Figure 3. When session agent SA1 receives successive packets of sequence number 1 and 2 from sender Sx, it caches the received packets, and then immediately returns acknowledgment (ACK) packets to the sender. After that, SA1 duplicates the received packets to transmit them to receivers Rx1, Rx2, and the session agent SA2. It keeps the cached packets for future retransmission until all packets sent are acknowledged. When it receives ACK packets from all receivers and the session agent, it discards the cached packet. Note that session agents perform flow control to avoid the overflow of an output buffer. It means that the output packets transmitted from session agents are not lost when forwarding the cached packets.

B. End-to-End control

In the end-to-end packet transmission control method, session agents collaborate to control the transmission rate of each partitioned TCP connection on an end-to-end basis. This method controls transmission rates with regard to the bottleneck link bandwidth between group members. Figure 4(b) illustrates the operation of session agents in this method. Similar to Figure 4(a), this method also provides packet caching, duplicating, and forwarding functions. The difference from the node-to-node control method is that the session agent returns ACK packets to the sender after receiving it from all the receivers and session agents. Consequently, session agents discard cached packets before they receive new data packets, and this suppresses the packet cache size of session agents. Note that the output packets transmitted from session agents are not lost, as described in section III-A.



(a) Node-to-Node packet transmission control method



(b) End-to-End packet transmission control method

Figure 4. Operation of session agent

IV. SIMULATION ENVIRONMENT

We evaluate the performance of the proposed methods through computer simulation in contrast with the unicast-based group communication method (traditional method). We used the network simulator ns-2.31, after adding the functions of the proposed methods.

A. Simulation model

Figure 5 shows the network topology used in this simulation. In this model, all end nodes which be discretely located in four networks *a, b, c* and *d* form a group and communicate with each other. Nodes m*a1*, m*bx*, m*cx*, and m*dx* (x = 1–5) represent group members. *a, b, c*, and *d* represent network identifiers and *x* represents the member identifier on each network. In one-to-many communication, node m*a1* sends data to all receivers through session agents SA1 and SA2, while nodes m*a1*, m*b1*, m*c1*, and m*d1* send data to members whose member identifier is *1* in the many-to-many communication. Note that all session agents operate as normal routers when the traditional method is employed.

Table I summarizes the simulation parameters. The propagation delay time of the link between nodes SA1 and SA2 varies from 10 to 100 ms and that of other links is set to 10 ms. The bandwidth of all links is set to 100 Mb/s. In addition, the number of members on each network varies

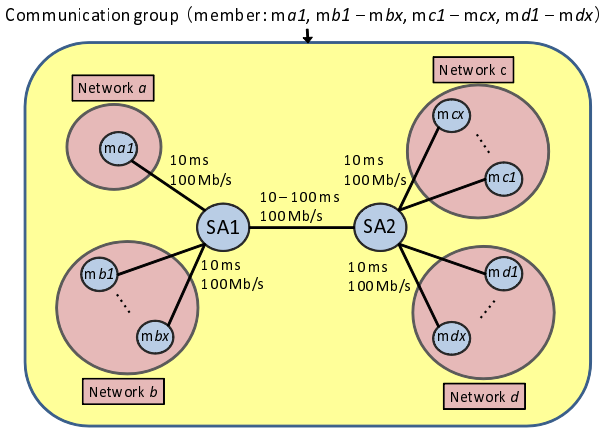


Figure 5. Simulation model

Table I
SIMULATION PARAMETERS

The number of group members on each network	1–5
Output buffer size of router	200 packet
Buffer size of SA for packet caching	∞
Data size	5 MB
Transport protocol	TCP SACK
Packet size	1500 B
Simulation time	100 s

from 1 to 5. To investigate the required cache size of session agents, the buffer size of the session agent for packet caching is set to infinity. The output buffer size of each node is set to 200 packets. All senders employ TCP SACK for data transmission, and the TCP packet size is set to 1500 Bytes.

After the simulation starts, the sender repeatedly transmits data with a size of 5 MB, based on the following traffic generation model, to vary the traffic load. In our simulation, the traffic load is defined as $t_r/(t_r + t_i)$, where t_r [s] and t_i [s] represent the required time for data transmission and the idle time, respectively. Each send time of the sender is $(t_r+t_i)(x+i)$ [s], where x and i represent a uniform random variable ranging from -0.5 to 0.5 and an incremental variable, respectively. We conduct simulation experiments for 100 seconds each and show the performance of each method by the mean value of five simulation runs.

B. Evaluation indices

To evaluate the effectiveness of the proposed methods, we focus on bandwidth consumption, reception throughput, effective throughput, and the cache ratio of session agents as evaluation indices. In group-oriented communications, the amount of data that all group members are correctly received is important. Thus, the reception throughput is defined as the minimum value of the reception throughput of each member. To evaluate the transmission efficiency, the effective throughput is defined as T_r/D, where T_r and D represent the reception throughput and the transmission rate of the sender, according to each traffic load, respectively. To investigate cache efficiency, the cache ratio is defined as C_{max}/D_{max}, where C_{max} and D_{max} represent the maximum cache data size and transmitted data size of all senders, respectively.

V. SIMULATION RESULT

In this section, we discuss the effectiveness of the proposed methods compared with the traditional method.

A. One-to-many communications

Figure 6 shows the bandwidth consumption of a link between nodes m*a1* and SA1 when the number of members in each network varies from 1 to 5. In this figure, "Unicast," "End-End," and "Node-Node" represent the results of the traditional method, the end-to-end, and the node-to-node packet transmission control methods, respectively. The bandwidth consumption of the proposed methods is lower than that of the traditional method because the traditional method requires multiple connections in the link in proportion to the number of members, whereas both proposed methods can aggregate these connections into one connection. Therefore, both proposed methods consume only the required bandwidth for data transmission and improve network efficiency.

Next, Figure 7 shows the reception throughput when the propagation delay time of a link between nodes SA1 and SA2 is 10 and 100 ms. The reception throughput of the proposed methods is higher than that of the traditional method. In the traditional method, a member whose round trip time (RTT) is longer than that of other members attains lower throughput performance due to the competition among multiple connections. Especially, when the traffic load is high, the short RTT member occupies most of an available bandwidth. Therefore, the reception throughput which means a minimum throughput among members reaches zero as the traffic load increases. On the other hand, both proposed methods achieve a high reception throughput in proportion to the traffic load because of the avoidance of competing with multiple connections. Moreover, both proposed methods decrease the impact of the propagation delay time due to the efficient utilization of the link bandwidth. Therefore, both proposed methods improve throughput performance.

Finally, Figure 8 shows the effective throughput. In the traditional method, the effective throughput decreases in proportion to the traffic load due to degradation of the reception throughput, while the proposed methods maintain a high effective throughput regardless of the traffic load. These results show that, the proposed methods improve network efficiency and communication performance.

B. Many-to-many communications

In this section, we discuss the results of the case where m*a1*, m*b1*, m*c1*, and m*d1* are senders. Figure 9 shows the average bandwidth consumption of four links between

 ISBN:978-1-61208-113-7

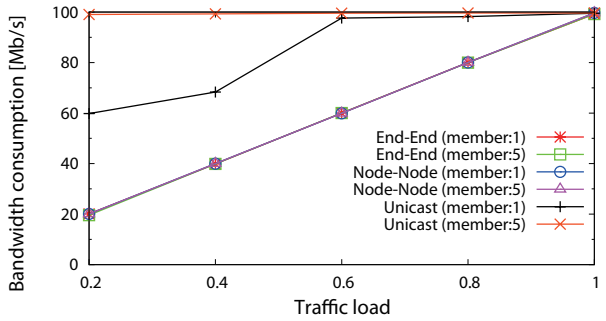


Figure 6. Bandwidth consumption of a link between ma*1* and SA1

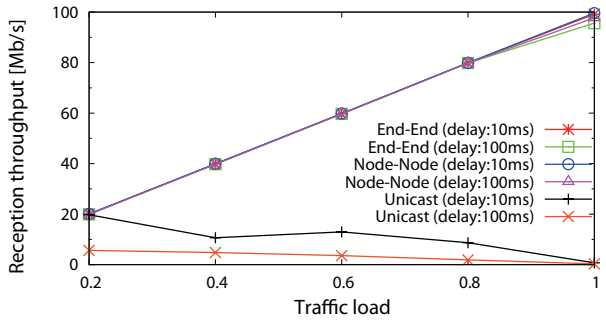


Figure 7. Reception throughput

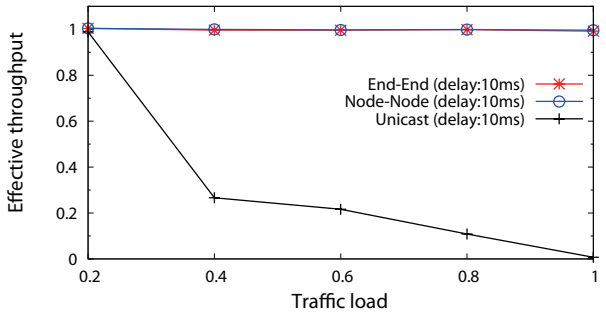


Figure 8. Effective throughput

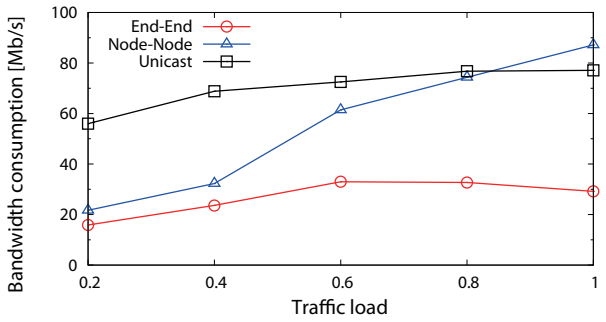


Figure 9. Average bandwidth consumption of links between senders and SAs

each sender and the SA. When the traffic load is low, the bandwidth consumption of both proposed methods is lower than that of the traditional method because each sender in the traditional method must transmit the same data to three members, whereas both proposed methods eliminate redundant data transmissions. On the other hand, when the traffic load is high, the bandwidth consumption of the node-to-node packet transmission control method is higher than that of the other methods. In the node-to-node packet transmission control method, each sender can optimize the available link bandwidth, so the bandwidth consumption increases in proportion to the traffic load. In the end-to-end packet transmission control method, competition among multiple flows occurs, so that the throughput of each flow from all senders is degraded, which decreases the bandwidth consumption.

Next, Figure 10 shows the average reception throughput of four receivers. The node-to-node packet transmission control method and the traditional method degrade the reception throughput in proportion to the traffic load. This is because the traditional method decreases the transmission rate of all members due to the competition among multiple connections. In the node-to-node packet transmission control method, the transmission rate of each sender increases in proportion to the traffic load, leading to an enormous increase in the number of ACK packets sent from session agents to each member. As a consequence, the reception throughput decreases drastically due to the obstacle of data transmission that is caused by the explosion in ACK packet transmission from session agents. On the other hand, in the end-to-end packet transmission control method, the transmission rate of each sender is lower than that of the node-to-node packet transmission control method due to the competition among multiple connections. As a result, each session agent can send more data packets to each receiver compared with the node-to-node packet transmission control method—it avoids the high level of ACK packet forwarding. Therefore, the end-to-end packet transmission control method provides high performance throughput.

Finally, Figure 11 shows the average effective throughput of the four senders. When the traffic load is high, the effective throughput of the end-to-end packet transmission control method is higher than that of other methods. From the above results, we can see that the end-to-end packet transmission control method improves network resource usage and throughput performance, even under high traffic loads.

C. Cache ratio of session agent

In this section, we investigate the required cache size of session agents. Figure 12 shows the cache ratio of SA1 when the sender is m*a1* only and m*a1*, m*b1*, m*c1*, and m*d1* are senders. When the sender is m*a1* only, the cache ratio is less than 10%. This is because the difference between the received and transmitting rate of session agents is small. On the other hand, when four group members are senders,

 ISBN:978-1-61208-113-7

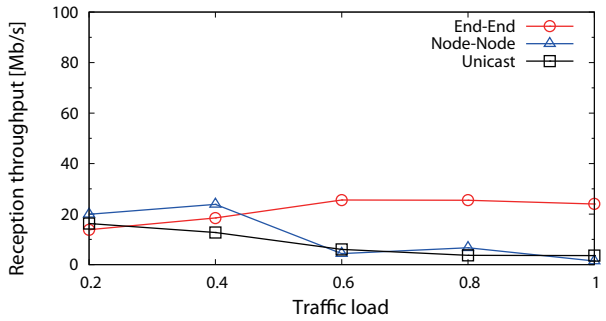


Figure 10. Average reception throughput

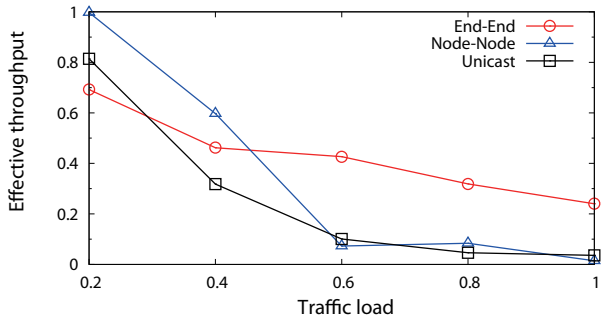


Figure 11. Average effective throughput

the cache ratio is higher compared with the case when m*a1* is the only sender. When four group members are senders, the difference between the received and transmitting rate of the session agents becomes large due to the competition among multiple flows. Moreover, the cache ratio reaches about 80% in the node-to-node packet transmission control method. In this method, the session agent requests new data packets to the sender as soon as it receives data packets, so that it receives a large number of data packets before discarding cached packets. On the other hand, in the end-to-end packet transmission control method, the session agent requests new data packets to the sender after receiving ACK packets from all receivers. Therefore, this method suppresses the packet cache ratio. These simulation results show that the end-to-end packet transmission control method improves cache efficiency.

VI. Concluding Remarks

We focus on the efficient group-oriented communication based on the split connection methods, which introduced a session agent with a packet caching function. In this paper, we proposed packet transmission control methods to improve cache efficiency of session agents, especially under high traffic loads. From our performance evaluations, the proposed methods can improve throughput performance and cache efficiency. In particular, the end-to-end packet transmission control method achieves excellent performance even under high traffic loads. In future work, we will discuss

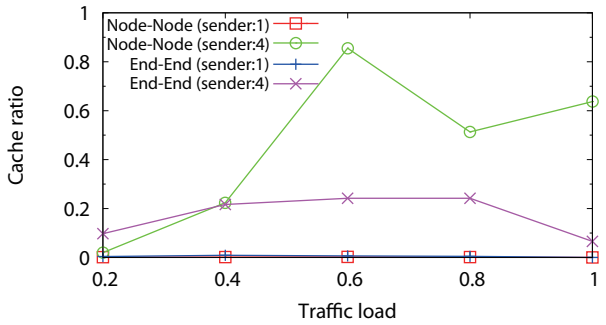


Figure 12. Cache ratio of SA1

an algorithm to find the optimal location of session agents.

Acknowledgment

This work was supported in part by the National Institute of Information and Communications Technology, Japan and the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (S) (No. 18100001).

References

[1] Y. Takahashi, K. Sugiyama, H. Ohsaki, T. Yagi, J. Murayama, and M. Imase, "Group-oriented communication: Concept and network architecture," *Proc. First International Workshop on Security of Computer Communications and Networks (SoCCaN2008)*, pp. 649–655, August 2008.

[2] H. Emina, H. Koga, M. Shimamura, and T. Ikenaga, "Split connection scheme for session-based group communication," *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim2009)*, August 2009.

[3] S. Y. Shi and J. S. Turner, "Multicast routing and bandwidth dimensioning in overlay networks," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 8, pp. 1444–1455, October 2002.

[4] M. Kobayashi, H. Nakayama, N. Ansari, and N. Kato, "Robust and efficient stream delivery for application layer multicasting in heterogeneous networks," *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 166–176, January 2009.

[5] W. Wang, D. A. Helder, S. Jamin, and L. Zhang, "Overlay optimizations for end-host multicast," *Proc. ACM International Workshop on Networked Group Communication(NGC2002)*, October 2002.

[6] L. Lao, J. H. Cui, M. Gerla, and S. Chen, "A scalable overlay multicast architecture for large-scale applications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 4, pp. 449–459, April 2007.

[7] K. To and J. Y. Lee, "Parallel overlays for high data-rate multicast data transfer," *Elsevier Computer Networks*, vol. 51, no. 1, pp. 31–42, January 2007.

[8] N. Ryoki, H. Koga, K. Kawahara, and Y. Oie, "Proposal of new layer of indirection between application and transport layers for flexible communications in IP networks," *Proc. IEEE/IPSJ International Symposium on Applications and the Internet (SAINT2008)*, pp. 289–292, July/August 2008.

 ISBN:978-1-61208-113-7

Efficient Location-aware Replication Scheme for Reliable Group Communication Applications

Yuehua Wang[1,2,3], Zhong Zhou[1,2], Ling Liu[3], Wei Wu[1,2]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
[2]School of Computer Science and Engineering, Beihang University, China
[3]College of Computing,Georgia Institute of Technology, USA
yuehua.wang1981@gmail.com,{zz,wuwei}@vrlab.buaa.edu.cn
lingliu@cc.gatech.edu

***Abstract*—In this paper, we present an efficient replication scheme designed to provide scalable and reliable group communication services for end system nodes that are widely dispersed in the network. This scheme is highlighted with three features. First, it employs both remote and neighbor replicas and intelligently avoids the services from interruptions while keeping replication cost low. Second, by using a tunable parameter, this scheme enables the nodes to adaptively make the most promising replica placement decision according to their specific network statuses. Third, a novel recovery technique is presented to minimize service recovery delay. The experimental results show that the proposed scheme is highly efficient in improving service reliability compared with existing neighbor-based replication scheme, random replication scheme and simple alternative schemes. The service recovery delay is greatly reduced by combining with the proposed recovery scheme.**

Keywords-location; replication; reliable; flexible; group communication;

I. INTRODUCTION

Overlay network has emerged as a promising paradigm to support group communication applications such as Video Conference, IPTV [1], Distributed Interactive Simulation, Local-based Notification [2] and Really Simple Syndication, which are characterized by exchanging contents among multiple end system nodes. It provides high scalability and availability for applications by harnessing loosely coupled and inherently unreliable end system nodes at the edge of the Internet.

However, one of the main challenges in overlay network currently facing is how to build reliable group communication services based on that. Due to the unreliable nature of nodes, it is necessary to find an effective way to deal with unpredictable node departures and nondeterministic network partitions so that it can provide stable and reliable support for the applications of group communication. One approach to address this problem is to create and maintain a set of node copies to mask failures from applications in the presence of node failures and network partitions. While data replication has been well-studied in the context of both unstructured and structured networks, they are often costly or perform inefficiently in dealing with the failures when facing a high churn rate.

In this paper, we develop an efficient replication scheme that is dedicatedly designed to provide scalable and reliable group communication applications such as Multi-party Online Game, News Feed Propagation, Instant Messaging, Popular Internet Radio Service, and so on.

This scheme has three features. First, it uses both remote and neighbor replicas and intelligently avoids the services from interruptions while keeping replication cost low. Second, this scheme provides the nodes with a large degree of flexibility in deciding replica placement. By using a tunable parameter, each node can adaptively make its replica placement decision according to its specific network statuses. Third, it provides a novel technique to minimize service recovery delay. This technique is derived from independence of service recovery and connectivity recovery.

To investigate the performance of the proposed schemes, they are applied to GeoCast [3], a geographically distributed overlay system. The experimental results show that compared to existing replication schemes [4–7] our scheme is highly efficient in improving the service reliability and the service interruption time is greatly reduced by combining with the proposed recovery scheme.

The rest of this paper is organized as follows. In Section 2, existing studies related to our work are discussed. Section 3 gives an overview of GeoCast. The design of our replica placement algorithm is presented and the setting of replication parameter is discussed in Section 4. In Section 5, extensive simulations and performance evaluations are presented. Finally, Section 6 concludes the paper and discusses future research directions.

II. RELATED WORK

A large number of research proposals [8–12] have been well designed to address the reliable content delivery problem by using overlay. In particular, a proactive fault-tolerant multicast routing protocol [11] is proposed to use backup paths to deal with link or path failures. Fei *et al.* [12] devised a dual-tree scheme to improve the resilience of multicast

 ISBN:978-1-61208-113-7

dissemination by constructing a secondary tree in addition to the primary tree normally used. Once failure occurs, it activates the path in the secondary tree.

PRM [8] is a multicast data recovery scheme proposed to achieve high delivery ratio in the presence of node/link failures. In PRM, each node forwards data to both its child nodes and a constant number of random nodes during content dissemination. Once failure occurs, those data duplications can be used to feed the orphans isolated by network partitions. However, the volume of extra data duplications and traffic might be significant for the applications like on-demand news propagation and multi-party online game.

Overcast [10] is motivated to provide service for bandwidth-intensive content dissemination. In overcast, nodes perform content caching and failure recovery operations to alleviate the influence of node or link failures. Starting with the root, some number of nodes is configured linearly, that is, each has only one child. The drawback of this technique is the increased latency of content distribution as the data has to traverse all those extra nodes before reaching the rest of overcast nodes.

Scattercast [9] proposed a infrastructure-supported architecture for scalable and reliable multicast support. In Scattercast, the multicast problems are partitioned into a set of smaller and simpler sub-problems that can be easily addressed with local knowledge of nodes in the same region. However, such divide-and-conquer approach might be expensive, especially in a network with high link and node failure rate. A large number of additional messages may be generated for the service recovery and new region formation when failures happen.

Our work differs from all other schemes by providing a proactive component used to augment the performance of application protocols. Instead of changing the infrastructures of the current applications as well systems, it is employed to provide reliable services to them while in a scalable fashion. In this paper, this is achieved by utilizing both remote replica nodes and neighbor replica nodes. We will present the scheme in details in the following.

III. Background

In this section, we describe architecture of GeoCast and introduce multicast mechanism on which the algorithm of this paper is based.

A. Architecture

GeoCast is a structured geographical proximity-aware overlay network, consisting of nodes equipped with GeoCast middle-ware. In GeoCast, nodes are equivalent in functionalities, each of which can perform: message publishing, message subscribing, message routing or all of them. Each node keeps a set of information about other nodes in the network in its $peernodelist$. Such list contains two kinds of nodes: immediate neighbor node and shortcut node. Similar to CAN, two nodes are considered to be immediate neighbors when their intersection is a line segment. The term *shortcut node* refers to the node which is *old neighbor node* for a given node. As nodes arrive or depart, neighbor nodes may become shortcut nodes when they are not adjacent to the given node. Instead of removing old neighbors from the lists, GeoCast keeps those nodes in $peernodelist$s and uses them to speed up the procedure of message delivery.

B. Multicast Service

In GeoCast, multicast service is introduced as one component for supporting group communication between nodes. It has two benefits. First, it releases high link stress caused by message transitions among nodes in group communication session and consequently improves network resource utilization. Second, it reduces the delay of the message delivery from the publishers to the subscribers. Instead of transversing the data from the publishers, the subscribers often can get it from their parent nodes within a short delay. For each session, there exists a spanning tree that is an acyclic overlay connecting all the participants of the session. It is used by the publisher node for content dissemination. Detailed algorithms and examples of the multicast service establishment and maintenance process with node arrivals and departures are presented in [3].

IV. Location-aware Replication Scheme

To provide reliable group communication services, a location-aware replication scheme is designed. It takes advantages of both the neighbor node and the remote node to improve the reliability of the services offered by the unreliable nodes while keeping the overhead low. The idea is derived from that the random replica nodes can always provide high reliable services for the nodes in the presence of failures mentioned in [6, 7]. However, it comes at high cost of replica detection and maintenance. To avoid that, in our scheme, only the nodes in the $peernodelist$s are considered in replica placement. The benefits are two-fold. First, it reduces the replication cost by employing the nodes in vicinity. Second, it alleviates the influence of network partitions on the service quality by deploying the data copies on shortcut nodes.

In this section, we first introduce patterns of failures and then present details of our replication scheme.

A. Failure Pattern

We consider a network consisting of a number of unreliable nodes. At any point in time, nodes can become unavailable for various reasons such as node departure, computer crash, improper program termination and traffic congestion. Note that similar operations will be performed to deal with node departure and failure in GeoCast, we use the term failure to mean either departure or improper failure.

 ISBN:978-1-61208-113-7

Based on the distribution of node failures, two failure patterns are identified: distributed failure and centralized failure (i.e., network partition). In the pattern of distributed failure, node failures are scattered over network. There always exist available end nodes around single failed node that can detect and repair its failure. In the centralized failure pattern, some nodes of the network appear to be unreachable from certain nodes but not others. Once it happens, the entire network might be partitioned into multiple, isolated overlay network parts. In this case, if multicast services continue to operate on this disconnected network, it might be interrupted and have nothing to do but wait for self-recovering from network partition. To provide the best quality of service, it is necessary for replication schemes of applications to deal with the failures of different patterns while keeping cost low. Inspired by this, our scheme is proposed and we will present the details of our scheme in the next section.

B. Replication Scheme

This section focuses on the two main components of the replication scheme design. First, the parameter used to leverage the benefits of neighbor nodes and shortcut nodes is defined. Second, we specify how the replication scheme is performed in a network of nodes with heterogeneous capacities.

1) Parameter Definition: It is desirable for nodes to maintain both the neighbor replica node and the shortcut replica node to improve replication efficiency in terms of both reliability and scalability. In our design, a tunable parameter α is introduced to adjust the importance of those nodes. The parameter value ranges from 0 to 1. The smaller it is, the less important the shortcut nodes are, the more likely it leads to a higher replication cost than that of scheme employing nodes residing in adjacent and vice-versa.

Interestingly, looking for the best values for α is essentially finding the best tradeoff between reliability and scalability. Consider a node p who has a replica placement task with replication degree r. When increasing α, node p reduces the number of neighbor replica nodes which causes the probability of service remaining available in the presence of network partitions to increase. However, larger α also leads to more shortcut replica nodes which in turn increases replica creation and maintenance cost. Based on those facts, we find the α value can not be too high or too small for the purpose of minimizing the replication cost and optimizing the reliability of service. Given an average of $O(2d)$ immediate neighbors is kept on each node in GeoCast [3], we have: for any node p,

$$\alpha = \begin{cases} 0.5, & r \leq 2 \\ \frac{1}{r}, & 2 < r \leq \tau \\ \frac{r-\tau}{r}, & r > \tau \end{cases} \quad (1)$$

where τ is the number of neighbor nodes in the $peernodelist$ of node p. The motive behind that is to relatively reduce the importance of shortcut replica so as to minimize the link stress imposed by placing and maintaining shortcut replicas. Typically, for a given τ, as increasing r, the number of neighbor replica nodes also increases. After r reaches $\tau + 1$, neighbor replica number reaches its upper bound τ, that is, no neighbor node is left in the $peernodelist$ that can be selected as the replica node. In this case, more shortcut nodes tend to be involved in the replica placement.

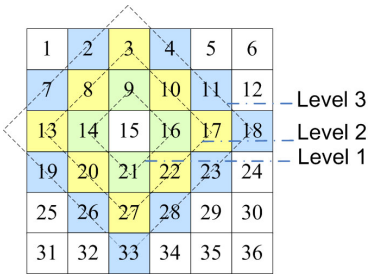


Figure 1. Relationship among nodes

2) Replica Placement: Each node in the multicast sessions deploys r nodes with available capacity in the network as replica nodes by performing the following operations.

Neighbor Selection The objective of this operation is to create neighbor replica nodes for nodes. Since the nodes' $peernodelist$ may or may not have "enough" available neighbor nodes, we introduce a new notation of *k-hop neighbor* B^k and define it as:

$$B^k_{E_i} = \begin{cases} \{immediate\ neighbors\ of\ E_i\} & k = 1 \\ \cup_{j=1}^{m_1} B_{B^1_{E_i}(j)} - B^1_{E_i} & k = 2 \\ \cdots & \cdots \\ \cup_{j=1}^{m_{l-1}} B_{B^{l-1}_{E_i}(j)} - B^{l-1}_{E_i} & k = l \end{cases} \quad (2)$$

where E_i denotes a node in the network and m_{l-1} is the number of nodes in $B^{l-1}_{E_i}$. When k=1, all immediate neighbors of node E_i are 1-hop neighbors of E_i. As mentioned earlier, two nodes are considered as immediate neighbors when the intersection of their regions is a line segment. If k=2, $B^2_{E_i}$ consists of the nodes that are the immediate neighbors of $B^1_{E_i}$'s entry nodes. For consistency, it requires the nodes in $B^1_{E_i}$ are not in $B^2_{E_i}$.

Fig. 1 gives a simple example of network to illustrate the relationship between nodes. In Fig. 1, node E_9, E_{16}, E_{21} and E_{14} are the immediate neighbors of node E_{15}. Nodes E_2, E_{10}, E_{17}, E_{22}, E_{27}, E_{20}, E_{13} and E_8 are immediate neighbors of $S^1_{E_{15}}$ that are included in $B^2_{E_{15}}$.

Neighbor selection starts with $B^1_{E_i}$. Node E_i first looks through its B^1 to see if there are $(1-\alpha)r$ nodes with available capacities. If so, node E_i adds them into its *replicalist* and replicats its multicast information on them, where *replicalist* is a list created for reordering the information of nodes which are selected as the replica nodes. Then this selection procedure terminates. Otherwise, E_i first adds all capable nodes at the first level into the *replicalist* and increases k by 1. Then the neighbor selection is performed at the second level. This procedure is executed repeatedly until either $(1-\alpha)r$ neighbor replica nodes are fetched or

 ISBN:978-1-61208-113-7

k reaches its maximum K. The parameter K is a system constant which is configured by default. The purpose of introducing K is to limit the cost of replica selection within a certain level. As suggested by [13], we set it to 2 in order to limit the replication searching cost within $O(4nd^2-2nd)$, where n is the network size.

Shortcut Selection The shortcut selection is performed in a similar manner as the first phrase. It starts with the set $S^1_{E_i} = peernodelist_{E_i} - B^1$, where $peernodelist_{E_i}$ is the*peernodelist* of node E_i. Then this procedure is continuously executed at the next level as there is no enough capable shortcut nodes. Similar to the definition $B^k_{E_i}$, $S^k_{E_i}$ is defined as $S^k_{E_i} = \cup_{j=1}^{m_{k-1}} B_{S^{k-1}_{E_i}(j)} - S^{k-1}_{E_i}$. This selection procedure is terminated when either αr shortcut replica nodes are fetched or k reaches its maximum K. Fig. 2 gives an example to illustrate the shortcut selection with setting of r=8 and α= 0.25. Node H first looks through $S^1_H = \{E_1, E_2, E_4, E_{14}, E_{18}\}$ and checks if there exists 2 shortcut nodes with available capacities. If so, no selection is necessary as enough replica nodes are detected. However, due to the heavy loads of nodes E_1, E_2, E_{14}, E_{18}, only E_4 is selected at the first level. To meet the requirement of $\alpha r = 2$, node H extends its search region and E_3 resided at the second level is then included in the *replicalist*. Similarly, capable nodes E_6, E_7, E_9, E_{11}, E_{12}, E_{16} that are marked with green circle are selected and then added into the *replicalist* after neighbor selection.

C. Replica Management and Failure Recovery

To avoid introducing additional overheads, the maintenance information of replicas are appended to the existing heartbeat message in our scheme. Every T seconds, heartbeat mechanism is performed, as well as replica maintenance. The parameter T is a constant that refers to the parameterized heartbeat period. We do not eagerly notify replicas to update theirs backup information if the relationship of multicast tree has been changed. It would be not done until heartbeat message between them is issued. In our scheme, multicast messages also serve as implicit heartbeat messages avoiding the need for explicit heartbeat messages. Long-time absence of heartbeat indicates that the node is gone and its corresponding region becomes an *island*, and thus boots the recovery process mentioned in [14].

Once a replica failure is detected, the update of $replicalist$ is triggered and one new replica node is selected by using the replica placement algorithm. In the procedure of replicalist updating, only the failed node is removed and replaced with a node in the set $B^k \cup S^k$. The new replica may or may not be a neighbor node of the host node, which relates to the character of the failed replica node and status of the nodes in the set $B^k \cup S^k$.

A node's failure triggers the recovery process. In terms of nodes functionalities in the network, there are two distinct recovery actions:

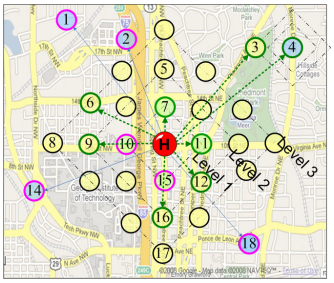


Figure 2. Shortcut Selection

- **Path Recovery** The goal of this action is to restore the services from interruptions when the failures of nodes in the multicast session happen. Rather than rebuilding the routing paths from the leaf nodes to the root nodes, the multicast information stored on the replica nodes are used for service recovery. It is done by replacing the failed node with one replica node and reconnecting with the parent and child nodes of the failed node. In our scheme, the node with high available capacity owns higher preference to be selected as the alternative node.
- **Region Recovery** This action is to search an alive node to take over the *island*. Rather than searching for leaf node of the *island*, our scheme tries to merge it locally with the help of failed node's neighbors. This process is stated by broadcasting merge request to all failed node's neighbors. Those nodes then check their regions to see if two of them can be merged. If more than one pair of siblings is around *island*, the pair with highest capacity is elected to handle such failure.

D. Replica selection policy

It is worth noting that there is wide variability in the available capacity of end nodes in the overlay network. Each node in the network has a fixed capacity. For different applications, it may refer to different factors such as processing power for responds, available storage space for data sharing services, or the available uploading network bandwidth for multimedia streaming applications. In our work, we use it to refer to the storage capacity of the nodes, denoted by c. It ranges from 0 to 1. c_i represents the maximum amount of storage that node E_i is willing to devote for serving other nodes. If $c_{E_i} = 1$, node E_i is overloaded; otherwise the node is under loaded. We use light and heavy to informally refer to nodes with low and high *utilization*, respectively. Periodically, each node calculates its load c and piggybacks it in heartbeat messages to update its status.

E. Analysis

In this work, one multicast service will be interrupted on a node only when all its replica holders fail in a short time interval. If it happens, a large number of re-subscription requests will be issued and routed among the nodes. To avoid this, it is a necessity for the replication schemes to employ enough replica nodes to deal with the failures of different

 ISBN:978-1-61208-113-7

pattern. Therefore, we next study the service reliability of a node archived by replication under an example setting to provide some insights into the performance of our scheme.

Consider a service s offered by node p. Let $R = \{e_1, e_2, ..., e_r\}$ be the $replicalist$ of p. Let P_i be the reliability of node e_i at time t_c. Then service reliability R_s can be calculated as:

$$R_s = 1 - \prod_{r}^{i=1}(1 - P_i) \quad (3)$$

$$P_i = P_i\{t > t_c | t > t_{s_i}\} = \frac{P_i\{t > t_c\}}{P_i\{t > t_{s_i}\}} = \frac{1 - P_i\{t \leq c\}}{1 - P_i\{t \leq s_i\}} \quad (4)$$

where R_s is the probability of at least one node in set R remaining in the network in next time slot. t_c and t_{s_i} refer to current time and arrival time when node e_i joins in the network. We use a Pareto distribution with shape parameter of 1.1 mentioned in [15] to estimate the distribution of node life time, represent by $F(x) = 1 - (1 + x/0.05)^{-1.1}$. As reported in [15], the average lifetime of peers remaining in the network is 0.5 hours. We have $R_s = 0.97$ with setting of r=4 and $Min\{P_i\} = 0.5$. It is clearly seen that the service reliability is greatly improved by the use of small number of replica nodes. This will also be confirmed by our experimental results in Section 5.

Our algorithm makes two effects to ensure scalability of the applications. One is to reduce the replica creation cost by intelligently using the information of the other nodes stored on the nodes. Rather than broadcasting its query request to number of nodes in the network, each node in the proposed scheme first inquires the nodes in its *peernodelist* for replica placement, which can improve the resource utilization. The other is to avoid high cost caused by remote-replica maintenance in message delay. In the procedure of replica placement, the nodes with available capacities residing in vicinity have high priority to be selected as the replica nodes. It enables either data updating or replicas' communication to be completed within a short delay.

V. Performance Evaluation

In this section, we conduct simulation experiments to evaluate the performance of the proposed replication scheme (called NR) and compare it against three schemes, the neighbor-based replication scheme (Neighbor) [4, 5], the random replication scheme (Random) [6, 7], and the hybrid replication scheme (Neighbor&Random) that is simply an extension of Neighbor and Random, which has the advantage of replication cost and reliability. The major difference in those schemes is in the replica creation.

A. Experimental Environment

We use Transit-Stub graph model from the GT-ITM topology generator to generate network topologies for our simulation. All Experiments in this paper run on 10 topologies with 8080 routers. Each topology consists of 8080 nodes with heterogeneous capacities. Similar to [16], the nodes in the network are assigned with various resource capacities: 5% nodes have 1000 units of capacity, 15% nodes possess 100 units of capacity, 30% nodes have 10 units of capacity, and the rest of nodes has 1 unit of capacity. The processing capacity of node is proportional to its resource capability. The more resource it has, the more powerful it is. Each unit of resource capacity allows nodes to maintain 10 files in theirs local memory.

Given that there is no linear relationship between the nodes' location and their associated link latency for delivering the message in IP network, we assign link latencies by following a uniform distribution on different ranges according to their types: [50ms, 80ms] for intra-transit domain links, [10ms, 20ms] for transit-stub links, and [1ms, 5ms] for intra-stub links. It allows messages transited between two nodes in the same domain have low network transmission delay. Nodes are randomly attached to the stub domain routers and organized into the GeoCast overlay network.

In each simulation, the multicast groups are built first. Both publisher nodes and subscriber nodes sequentially participate in the groups, following independent and identical uniform distribution. To simplify, we do not consider the case that the subscribers in the same group have different interests in publisher's content since it can be handled in a similar manner.

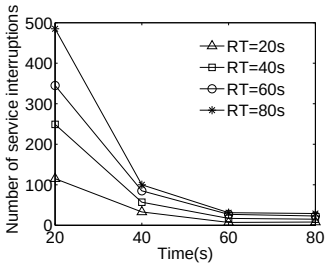
We run each trial of simulation for 60T simulated seconds, where T with setting of 120 is the parameterized heartbeat period. In our figures, each data point represents the average of measurements over 50 trials (5 trials in each topology with same setting in such a way the inaccuracy incurred by stochastic selection is minimized).

B. Results

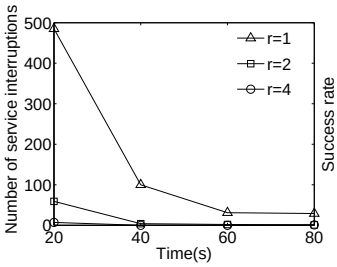
In this section, we investigate three main subjects using sets of experiments. We first evaluate how the efficiency of replication scheme in the presence of node failures. Then we study the effect of replication schemes on the system performance in terms of replication overhead. Third, the performance of the proposed recovery scheme is investigated.

1) Reliability: Similar to [17], we generate a sequence of node failures to study the effect of replication scheme on the service quality in terms of reliability. By following independent and identical exponentially distribution, the failure time of sequence nodes that are randomly chosen is assigned. As mentioned earlier, both the nodes and their replica holders may fail in a short time interval. If it happens, the multicast services of the nodes are interrupted and a large amount of re-subscription requests may be issued and transited on network for service recovery. To avoid this, high failure resiliency is desired in the current applications. As a matter of fact, it mainly depends on two factors, turnover

 ISBN:978-1-61208-113-7

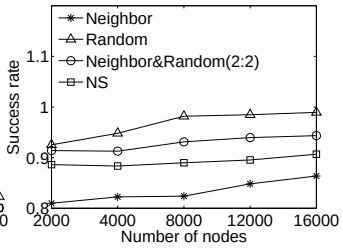


(a) r=1(FP=0.6)

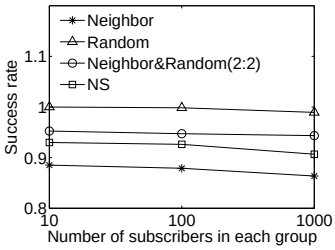


(b) r(FP=0.6 RT=80s)

Figure 3. Service interruption during runtime

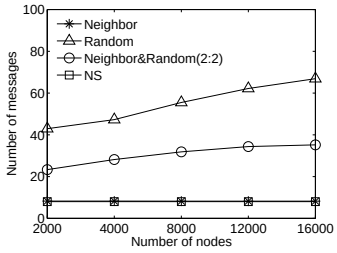


(c) different system size

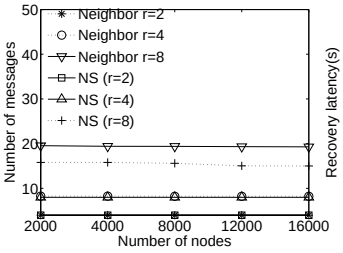


(d) different group size

Figure 4. Success rate(FP=0.6 RT=80s)



(a) different schemes



(b) different r

Figure 5. Overhead of replica creation

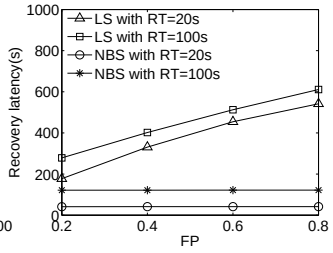


Figure 6. Effect of FP on region recovery latency

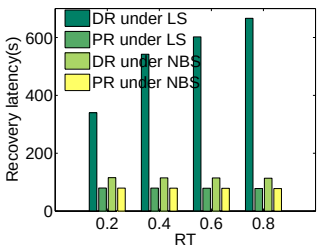


Figure 7. Effect of RT on recovery latency(FP=0.6)

time RT and replication degree r. RT refers to the time required for node initialization right after it is selected as replacement node to take over the *island* previously owned by the failed node. r refers to the number of replica nodes that are required to be placed by each node in the network.

To study the effect of those factors on the service interruptions, the following experiments are simulated in a set of 16000-node systems with FP=0.6. The value 16000 is the maximum number that we are able to simulate under constrains of the computer resources. Failure probability FP represents the faction of nodes failed at the run time. For instance, if FP is set to 60%, 9600 nodes will leave from the system at runtime. From results in Fig. 5, we observe that as time goes by, the number of service interruptions significantly drops. This is due to the fact that majority of service interruptions happen at the beginning stage of simulation. Meanwhile, it is important to notice that the metric become smaller when r is larger. When it increases to 4, the number of service interruptions is steadily below 10. Thus, we set r to 4 in the following discussion.

Fig. 4 shows the success rate as a function of system size. Success rate represents the ratio of the number of recoverable failures to the number of failures happened at runtime. We use the term *recoverable failure* to mean the service of a failed node that can be restored by one of its replica node after the failure happens.

For this and following plots wherein we vary the number of nodes N in the system, we use $N \in \{2000, 4000, 8000, 12000, 16000\}$. In Fig. 4, it is clearly shown that Random achieves a better performance than the other schemes. This is essentially because in Random replication, the replica nodes are likely to have higher probability to remain alive even facing high churn rates, which further confirms the conclusion mentioned in Section 4. Similar to Random, both Neighbor&Random and NS yield higher success rate than Neighbor. This can be explained by the fact that the nodes in Neighbor scheme are likely to replicate the information on their neighbor nodes. If the centralized failures happen, both the nodes and their replicas tend to fail simultaneously and thus it is hard to find one living replica node to restore the interrupted services.

2) Replica Creation Overhead: Fig. 5(a) shows the replication overhead as a function of system size. It is observed that both Neighbor and NS replication can successfully replicate the information on replicas with lower overheads than the others. This is due to the fact that the majority of replicas are discovered within 2 levels. Fig. 5(b) shows how the number of replicas affects the overheads of the replication schemes. As r increases, the overhead also increases. This is because more messages are issued from the nodes for replica selection when r is larger, which may lead to a poor application performance in scalability. To avoid that, NS enables the nodes to adaptively tune their parameters and deploy more shortcut replica nodes. The benefits are two folds. First, it reduces the replica searching cost. Second, it enhances the reliability of services in some sense. It indicts the efficiency of NS replication in overhead.

3) Recovery Latency: Recovery latency is the time interval between the time when a node detects a failure event and the time when new responsible node offers its services as the owner of *island*. To evaluate the performance of different recovery schemes in a dynamic environment, we vary failure probability from 20% to 80 % in 16,000-node network.

Fig. 6 shows the recovery latency as a function of fail-

 ISBN:978-1-61208-113-7

TABLE I
SUCCESS RATIO COMPARISON

	Hop	FP			
		0.2	0.4	0.6	0.8
NBS	1	86.70%	85.80%	85.70%	85.57%
	2	100%	100%	100%	99.99%
	3	100%	100%	100%	100%
	4	100%	100%	100%	100%
LS	Type	FP			
		0.2	0.4	0.6	0.8
	RF	60.75%	59.90%	59.60%	58.92%
	URF	39.25%	40.10%	40.40%	41.08%

ure probability FP. We observe that the proposed scheme (called NBS) steadily achieves better performance than leaf node scheme (LS) [14]. As FP increases, the performance improvement of our techniques is even more pronounced. It is obvious that LS scheme takes long time to recover the interrupted service than our scheme in terms of hop-count. This is due to the fact that some recovery are delayed by searching the alternative node [14], which is confirmed by the results in Table I. FR refers to the recoverable failure and UFR refers to the unrecoverable failure. In Table I, we can see that NR performs better than its competitor. All node failures can be recovered within 3 hops in all cases. For LS, when we set FP=0.8, 41.08% of the failures is unrecoverable at runtime. This is because that the failures can not be recovered until the alternative nodes are fetched in LS. In this case, the recovery latency increases exponentially, as shown in Fig. 6. The maximum latency of LS is around 60 times as many as that of NBS when FP = 0.8.

As a matter of fact, for both approaches, the recovery latency increases as increasing RT as shown in Fig.7. We observe that the scheme NBS beats its competitor in all cases and it is much more steady than LF in service recovery. Since both path recovery and region recovery conduct simultaneously, the recovery latency of NBS can be further minimized and is equal to $Max\{L_{RR}, L_{PR}\}$, where L_{RR} and L_{PR} refer to the latency of the region recovery and path recovery, respectively.

VI. CONCLUSION AND FUTURE WORK

We have presented a dynamic passive replication scheme. It is an extension of our previous work [3, 18]. In this paper, the proposed scheme is used to provide reliable group communication services for nodes who are unreliable. Through the simulations, the effectiveness of our scheme in different scenarios is demonstrated. The experimental results show that in the presence of node failures, the proposed scheme can efficiently recover services from interruptions within short recovery delay, which results in a better performance, compared to existing replication schemes for large scale group communication applications. For the future work, we would like to explore it in working systems and develop enhancement to make the proposed schemes more efficient in group communication applications.

VII. ACKNOWLEDGMENT

The first author conducted this work as a visiting PhD student at Georgia Institute of Technology. This work is partially supported by NSF NetSE, NSF CyberTrust, IBM SUR, Intel Research Council, 2008 China Next Generation Internet Application Demonstration sub-Project (No.CNGI2008-123) and Fundamental Research Funds for the Central Universities of China. The first author also thanks China Scholarship Council for supporting the visiting.

REFERENCES

[1] K. Kerpez, D. Waring, G. Lapiotis, J. Lyles, and R. Vaidyanathan, "IPTV Service Assurance," *IEEE communications magazine*, vol. 44, no. 9, p. 166, 2006.
[2] P. Persson, F. Espinoza, P. Fagerberg, A. Sandin, and R. Cöster, "GeoNotes: A Location-based Information System for Public Spaces," *Designing Information Spaces: The Social Navigation Approach*, pp. 151–173, 2002.
[3] Y. Wang, L. liu, C. Pu and G. Zhang, "GeoCast: An Efficient Overlay System for Multicast Application," *Tech. Rep., GIT-CERCS-09-14, Georgia Tech*, 2009.
[4] J. Zhang, L. Liu, C. Pu, and M. Ammar, "Reliable peer-to-peer end system multicasting through replication," in *P2P'04*.
[5] T. Chang and M. Ahamad, "Improving service performance through object replication in middleware: a peer-to-peer approach," in *P2P'05*, pp. 245–252.
[6] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, "Search and replication in unstructured peer-to-peer networks," in *Proceedings of the 16th international conference on Supercomputing*, 2002, pp. 84–95.
[7] Y. Liu, X. Liu, L. Xiao, L. Ni, and X. Zhang, "Location-aware topology matching in P2P systems," *INFOCOM 2004*, pp. 2220–2230.
[8] S. Banerjee, S. Lee, B. Bhattacharjee, and A. Srinivasan, "Resilient multicast using overlays," *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, no. 1, pp. 102–113, 2003.
[9] Y. Chawathe, "Scattercast: an adaptable broadcast distribution framework," *Multimedia Systems*, vol. 9, no. 1, pp. 104–118, 2003.
[10] J. Jannotti, D. Gifford, K. Johnson, M. Kaashoek, *et al.*, "Overcast: Reliable multicasting with an overlay network," in *Proc. Usenix Fourth Symp. Operating System Design and Implementation*, 2000.
[11] W. Jia, W. Zhao, D. Xuan, and G. Xu, "An efficient fault-tolerant multicast routing protocol withcore-based tree techniques," *IEEE Transactions on Parallel and Distributed Systems*, vol. 10, no. 10, pp. 984–1000, 1999.
[12] A. Fei, J. Cui, M. Gerla, and D. Cavendish, "A dual-tree scheme for fault-tolerant multicast," in *Proceedings of IEEE ICC*, 2001.
[13] V. Kalogeraki, D. Gunopulos, and D. Zeinalipour-Yazti, "A local search mechanism for peer-to-peer networks," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 307–314.
[14] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker, "A scalable content-addressable network," in *SIGCOMM'01*, vol. 31, no. 4, pp. 161–172.
[15] X. Wang, Z. Yao, and D. Loguinov, "Residual-Based Estimation of Peer and Link Lifetimes in P2P Networks," vol. 17, no. 3, 2009, pp. 726–739.
[16] J. Zhang, L. Liu, L. Ramaswamy, and C. Pu, "PeerCast: Churn-resilient end system multicast on heterogeneous overlay networks," *Journal of Network and Computer Applications*, vol. 31, no. 4, pp. 821–850, 2008.
[17] S. Saroiu, P. Gummadi, S. Gribble, *et al.*, "A measurement study of peer-to-peer file sharing systems," in *Proceedings of Multimedia Computing and Networking*, vol. 2002, p. 152.
[18] G. Zhang, L. Liu, S. Seshadri, B. Bamba, and Y. Wang, "Scalable and Reliable Location Services Through Decentralized Replication," in *International Conference on Web Services 2009*, pp. 632–638.

 ISBN:978-1-61208-113-7

I2P Data Communication System

Bassam Zantout and Ramzi A. Haraty
Department of Computer Science and Mathematics
Lebanese American University
Beirut, Lebanon 1102 2801
Email: rharaty@lau.edu.lb

***Abstract*---As communication becomes more and more an integral part of our day to day lives, the need to access information increases as well. Security is currently one of the most important factors to consider in our aim to achieve ubiquitous computing, and with it raises the problem of how to manipulate data while maintaining secrecy and integrity. This paper presents one of the most common widely used data communication systems for avoiding traffic analysis as well as assuring data integrity - I2P. I2P, just like every other technology aimed at securing data, has its pros and cons. The paper presents the benefits and drawbacks of I2P.**

Keywords - I2P; Traffic Analysis; Data Integrity.

I. INTRODUCTION

Since the day the Internet became a common and reliable mechanism for communication and data transfer, security officers and security enthusiasts rallied to enforce security standards on data transported over the globe. The goal was to achieve data integrity and confidentiality, while using a reliable data transport medium, which is the Internet. Whenever a user tries communicating with another recipient on the Internet, vital information is sent over different networks until the information is dropped, intercepted, or normally reaches the recipient. This information identifies where the request is coming from by revealing the user's IP and hence the geographical location, what the user needs from the recipient, and sometimes the identity of the user. The moment the recipient replies back, the same type of information is sent back along with a certain payload (meaningful content) for which the user had requested. Critical information traversing networks is usually encrypted. Sometimes encrypting the payload alone is not enough for users who wish to conceal their identities while communicating with recipients over the Internet. Take for example a reporter working undercover and sending critical information over the Internet to a country that is at war with where the reporter is residing in. If the reporter's identity is revealed, then the reporter might be convicted. Hence, concealing who is sending the information is sometimes much more important than revealing the information itself. In order to conceal the sender's identity, different implementations have proven successful. One of which is the invention of anonymous networks. Anonymous networks go beyond transferring information over the Internet, whereby theoretically, the implementations can be replicated on different communication technologies such as mobile devices, wireless networks, etc. [1, 2, 3]. In 2003, and due to a huge interest as well as considerable advances in P2P concepts, whereby numerous projects for distributed file sharing and P2P networking emerged, a new project called the I2P (*Invisible Internet Project*) was introduced to the public [4, 5, 6]. The main developers and supporters for this project remain anonymous to this date and call themselves nicknames whereby *jrandom* is the main developer and the person responsible for this project, which was later called I2P. *jrandom*, along with many developers studied different anonymous systems at that time (Tor, Tarzan, Freenet, Bitorrent) [7, 8, 9, 10] and then discovered and implemented new and unique ideas for distributed P2P anonymous systems, which promised better anonymity to its users [11, 12].

The paper discusses I2P and presents its benefits and drawbacks that will be discussed thoroughly including the newly introduced methodology. The remainder of this paper is organized as follows: Section two provides the background of I2P. Section three discusses I2P in detail and Section four provides a conclusion.

II. BACKGROUND

I2P is a low latency anonymizing mix network that offers its users a certain level of traffic analysis prevention; hence, hiding the identity of both the sender and receiver, while utilizing a large set of encryption standards to hide data content and to ensure payload delivery. Just like NetCamo [13], I2P is intended to be used with nodes that have I2P system installed. Moreover, just like Tor, I2P is capable of relaying traffic through multiple nodes using tunnels and encapsulated messages of data that are routed until the destination is reached. However, the key difference is that I2P is a message-based system instead of circuit-based as in Tor. Moreover, unlike Tor, I2P is a fully distributed system that does not rely on centralized directory servers to keep track of participating nodes and network performance. Instead, I2P utilizes a modified Kademlia algorithm [14] that handles network and node information that is distributed and maintained among different nodes in the I2P network. After several years of discussions and development, I2P is still considered in the alpha stage whereby the core components and driving engine have been

 ISBN:978-1-61208-113-7

changed frequently and will continue to change due to enhancements. I2P is still not concerned a fully reliable anonymous system, although developers and users can logon to the network for a test drive.

III. I2P AND GARLIC ROUTING

The following sections describe how I2P functions and what makes its corresponding components unique. It is important to note that I2P, while similar to Tor in some of its definitions, differs immensely in its design and implementation.

A. Garlic and Onions, Cells versus Cloves

The Second Generation Onion Routing Project, Tor, as well as the original onion routing design devised a system based on cells whereby a cell is of fixed size that contains encrypted information of either instructions to other onion router nodes, or data/payloads to be delivered to a certain recipient. The onion cell has fixed size in order to conceal hints about information or the content of the data being transmitted from and between nodes in the system. This fixed sized technique was considered as a security and anonymity enforcing mechanism since traffic analysis of fixed size cells could prevent against website fingerprinting, and target/sender correlation attacks. While this technique is indeed effective in high-end Tor nodes where millions of cells are passing every hour through these nodes; however, fixed cells prove inefficient when it comes to end-to-end attacks and time-based attacks since cells are not padded with random data and lack intentionally introduced delays (with latency considerations). Garlic routing was inspired from onion routing whereby garlic cloves are simply a combination of one or two onion cells in addition to extra padded information of random size. Hence, the atomic data unit for the I2P system is indeed the same as the onion system; however, not a single atomic unit is transmitted alone. Instead, previously encrypted onion cells are grouped together, with extra padding, as well as delay/no-delay instructions to other I2P nodes, and then packaged in so called Garlic cloves, which are then passed to other I2P nodes, in an encrypted format. The size of a clove as well as the number of onion cells differs between I2P nodes in order to add additional randomness to the system. As I2P nodes receive encrypted Garlic cloves, they are able to decrypt them (using public and private keys) and then treat each onion cell independently and sometimes with special latency and priority requirements sent by embedded instructions. Each I2P node is then able to repackage received onion cells using new encrypted garlic cloves and then send them to other I2P nodes.

This alone makes I2P a message-based system instead of a circuit-based system as in Tor. However, the notion of circuits and hops still exists. What is important to note is that if two users who have both installed the I2P client software, these users will be able to send information to each other in fully encrypted format, and thus prevent end-to-end attacks to non-global adversaries.

B. Tunnel and Communication amongst I2P Nodes

I2P utilizes a huge set of protocols, encryption standards, and P2P concepts in order to achieve the highest levels of anonymity for its users. This section describes I2P node and user communication in details. However, it is important to stop and visit two vital concepts in P2P networking and to I2P, which are DHT and Kademlia.

C. DHT (Distributed Hashed Tables)

A hash table is a simple algorithm that given a hash function and a certain input, then a *unique* output (depending on the hash function) is derived. Hash functions are extremely efficient in locating values that correspond to a certain input, and the notion of buckets is used to indicate many values a hash function could outputted to when a single input is used. A DHT is a similar concept for decentralized distributed systems whereby one is able to lookup information in a distributed system efficiently. Given two pairs of information *(name, value)* which are stored in a DHT, participating nodes can work together in maintaining and mapping these pair of information amongst each other with minimal amount of resource and network overhead. DHT was crafted after being inspired from inefficient distributed lookup services found in P2P implementations at the time. These P2P implementations where mainly focused at locating resources or files in distributed systems whereby three methodologies were used:

1) Centralized Indexed System: a central system was assigned whereby participating nodes pushed whatever resource listing they had to this system. The central node then performed indexing on this information and any user who wished to locate data present on the distributed nodes queried the central system for the location of data. A similar system that received its 15 minutes of fame was Napster that later faced enormous slowdowns as the size of the files and data increased.
2) Flooded Query System: is another system that required for each query, a user issued, to be distributed to all participating nodes in the system. Although this might reveal the most updated results since no central system performance or update delays might occur; however, allowing such a system to scale was improbable since as the number of search queries increases and as the number of nodes increases, the number of broadcasts and replies also increase. Gnutella is an example of such a P2P system.

 ISBN:978-1-61208-113-7

3) Heuristic Key Based Routing: was utilized by Freenet, whereby a resource was associated with a key and resources with similar keys are located in a cluster or group of similar nodes. So based on the key a user issues, a key based routing is implemented and the query is directed to these set of nodes instead of being broadcasted to all nodes, or a centralized system.

DHT was developed to overcome the above points and is characterized by the following:

1) Decentralization: each node is an independent node that does not rely on a centralized system for coordinating tasks and locating information.
2) Scalability: A DHT system is able to scale highly (millions of nodes) while keeping its phenomenal and efficient search capabilities as is.
3) Fault Tolerance: nodes in a DHT system may join and leave the system while keeping all stored information in the system intact.
4) Performance: Given ***n*** nodes in a system, then as the number of nodes increases or decreases, the system is able to retrieve information in the O(logn) (Big O notation).

DHTs received a great deal of attention by many academic institutes for which implementations like Chord, Kademlia, CAN, Pastry and Tapestry where developed. The implementations differ in some details; however, the overall concept is the same whereby three components are identified: *keyspace*, *keyspace partitioning*, and *overlay network*.

A *keyspace* is a set of sequence of bits of a certain fixed length ***F***. For any content that needs to be stored in a DHT system a filename is hashed and one obtains a hashed value of size ***F*** that corresponds to a certain resource ***R***. The data along with the hashed filename are introduced into the DHT network and the DHT system forwards (through the *overlay network*) such information amongst DHT nodes until the data arrives to the DHT node responsible for keeping track of this file information. A query then given to any node in the DHT system about this filename is hashed and then forwarded (using the *overlay network*) till it arrives to the node responsible for such information.

Nodes in a DHT system are conceptually arranged in circular ring network although physically nodes may be geographically dispersed over the globe. Each node is aware of its successor and predecessor, and traversal of the nodes usually occurs clockwise.

For two keys k1 and k2, *keyspace partitioning* is illustrated as the distance between k1 and k2 represented by $\delta(k1,k2)$, whereby the distance does not relate to network latency or geographical location. Each node in the DHT network is then give a key as its identifier; hence, a node with ID *i* owns all the keys for which *i* is closest to. The *keyspace* is split into contiguous segments whose endpoints are the node IDs, whereby for any two nodes *i1* and *i2* for a key *kx* then *i2* holds all the keys that fall between *i1* and *i2*.

The DHT system may introduce a pool of nodes whereby each of pool is responsible for replicated data content to cater for node joins and departures in the system as displayed in Figure 1.

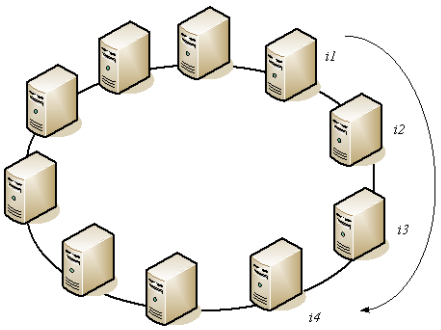


Figure 1. Sample DHT system.

The *overlay network* can be described by the fact that each node keeps track of its adjacent nodes; hence, traversing the system may require an O(n) (Big O notation. Although this might become extremely inefficient for large systems, a series of enhancements where introduced, as in Chord and Kademlia, which reduced the traversal to O(logn) by adding a routing table to each node (through continuous lookups as nodes join and leave the network or as queries traverse the network, nodes are able to keep track changes). The search method using the *overlay network* is executed using a greedy algorithm that forwards/routes searches to the closest node holding a key similar or close enough to the original key. Of course, different DHT designs have different implementations of the *overlay network*.

DHT is considered a core infrastructure that has been used to build many complex systems that have been widely adopted and later modified by many implementations like Bittorrent, I2P, Coral Content Distribution, eMule, and Freenet, and many others.

D. Kademlia

I2P designers have adopted a DHT implementation by Petar Maymounkov and David Mazi`eres, from the University of New York, called Kademlia. This DHT implementation is capable of running in a network where a lot of node joins and departures occur and hence uses a XOR-based metric topology whereby the distance between two nodes is computed as the XOR of the node IDs (knowing that the node IDs are in a certain increasing sequence). Additionally, queries about keys and nodes in a network are recorded by every DHT node through which a query traverses through. This, along with efficient data retrieval (O(logn)) due to a good routing implementation for

locating nodes and data in nodes, makes Kademlia a good candidate for any DHT implementation.

E. I2P Tunnel and Node Characteristics

The I2P network is composed of I2P routers that relay encrypted garlic cloves, and I2P users transmitting and receiving such cloves to each other. I2P routers and destinations (or end-user client nodes) have distinct identification through cryptographic identities, which enables them to send and receive messages as well as form encrypted tunnels. Each I2P node or router in the network has inbound and outbound tunnel(s) established and connected to other I2P gateway(s). The number of tunnels can be increased to form different routes by simply connecting to different I2P gateways. When a message needs to be relayed from a sender to a recipient the message goes through the senders outbound tunnel to the end-point of the tunnel, which is another I2P gateway and that gateway then forwards the message to through a series or hops (or directly) to the gateway of the recipient for which then the message traverses the recipients inbound tunnel and gets decrypted at the recipients node. Senders have no information about the path the message will take except for the gateway the senders have used to release the message. Each I2P router in the I2P network is able to add delays, introduce additional padding, and route information according to a node directory lookup called the *NetDB* (network database based on the Kademlia algorithm). Figure 2 better illustrates the I2P topology.

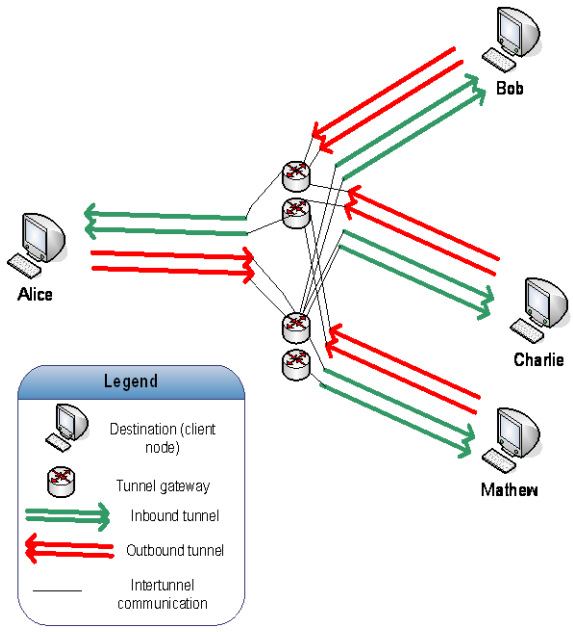


Figure 2. I2P network – A sample of inbound and outbound tunnels used for communication.

In the above illustration, Bob is able to send Alice information in a couple of steps by:

1) Querying the NetDB to retrieve information about routers identity and encryptions keys as well as destination's public keys and reachability of gateways and destinations. This information is stored in the NetDB under two categories: the *RouteInfo* and the *LeaseSet*.
2) Sending messages through its outbound tunnel to a router then in turn converts the stream into the inbound tunnel of Alice.
3) Getting back replies from Alice through Alice's outbound tunnel that in turn gets converted by I2P routers to Bob's inbound tunnels.

As such one can notice that I2P does not have any entry and exit nodes like Tor, and data is encrypted end-to-end among peers in the system. However, in order to achieve this, both sender and recipient need to be on the I2P network connected to at least a single I2P node with two tunnels (inbound and outbound). The amount of tunnel creations is user based. However, the default is creating four tunnels - two for inbound and another two for outbound communication respectively. This could be justified in the event that if one tunnel goes down for any reason, then a secondary tunnel exists. Moreover, adding more tunnels means having different routes to the same destination or a set of destinations. I2P users are also able to choose the number of hops found on the path to a certain destination whereby the illustration shows only a single hop from Bob to Alice. I2P currently supports a maximum of two hops before traffic is routed and delivered to its final destination. Tunnel creations are time based and change every 10 minutes in order avoid any types of attacks on the tunnel encryption. Even if tunnel encryption has been compromised, then payloads - packages in garlic cloves, have already a multilayer of synchronous and asynchronous encryption standards to assure data integrity and anonymity. Another type of tunnel exists, that has not been shown in the preceding graphical illustration, used for I2P node discovery. An explanation about node discovery is discussed in the next section.

Nodes in an I2P network are classified into four different categories according to speed and reliability:

1) **High Capacity**: are nodes that have a higher connection capacities and uptime than the average number of nodes connected to the system.
2) **Fast**: are nodes that are categorized as "high capacity"; however, they are considered with fast connection due to bandwidth (throughput) compared also to the average number of nodes with connection speeds.
3) **Not Failing**: are nodes if it is neither high capacity nor failing.

 ISBN:978-1-61208-113-7

4) **Failing**: are nodes that frequently join and leave the network, or that are queried but found to be always unavailable.

Although there is an additional node classification called "Well Integrated", which means that a node "integration calculation" is above average, there is no explicit information about what that really means. Since I2P is a decentralized and distributed system, with a large number frequent joins and departures of end-user nodes that form the majority of nodes in the system, then there is no way of determining exact node statistics similar to a centralized system as Tor. As a result, node information and categorization is usually kept at every node in the system with the help of Kademlia's node traversal and NetDB data querying. Hence, two nodes that are not well-integrated in the network or with network problems may have varied answers about a particular node in the network.

F. Peer Selection and Tunnel Creation

When a new user wishes to join the I2P anonymous network, the user downloads and installs the I2P software from the official website after performing a checksum on the software itself (to ensure data integrity and that the software has not been tampered with). The I2P software is a Java based client that enables users to look for an available I2P node and connect to that node and therefore join the I2P anonymous system. However, one might wonder how is this done when I2P itself is a distributed and decentralized anonymous network that is hard to keep track of, and connect to, especially when a new node tries to connect for the very first time. As a solution, the I2P developers have introduced a set of hosts' IPs that have a good uptime (*Fast* or *High Capacity*) and which are considered reliable hosts - once the I2P software launches it bootstraps using a randomly selected I2P host IP from the set of preconfigured (*Fast* or *High Capacity*) IPs and logs on the I2P network. If bootstrapping is unsuccessful then the client software can choose another IP until a connection occurs. Once connected, a series of node investigations is carried out by the newly joined node using a second type of tunneling used only to traverse the I2P network looking for available nodes. Node traversal begins taking place gradually as the new node starts building tunnels with random nodes in the system (Kademlia DHT algorithm). At every tunnel creation the new node will query existing I2P nodes for available nodes for which it can connect to, and since tunnels do not usually last long, the rate of discovery becomes high and the list becomes larger with a considerable uptime.

NetDB stores two sets of data: *RouterInfo* and *LeaseSets*. *RouterInfo* gives users or nodes in the system the set of information to contact a specific router in the network. *LeaseSets* gives nodes in the system information for contacting a particular destination node where data will be delivered, also composed of destination tunnel address(s), public keys, and the destination's tunnel uptime for path reliability. While *RouterInfo* information may not change frequently since users connect to dedicated gateways, *LeaseSets* for nodes do change frequently since users often change tunnels every 10 minutes and hence reachability is changed too. End user nodes usually choose first *Fast* or *High Capacity* I2P nodes so that tunnel creation is reliable for sending and receiving data, then the nodes establish another set of inbound and outbound tunnels called the *exploratory tunnels* with less capacity (*Not Failing*) nodes to query the network about other available nodes and to obtain the NetDB information carrying encryption specifications about how to connect and how to reach other nodes.

G. System Encryptions Standards for Communication

I2P uses four different types of cryptographic algorithms for ensuring communication reliability, anonymity, and data integrity during data transmission through multiple paths. As a result, symmetric, asymmetric, signing and hashing algorithms have been used all together to strengthen the security on communication. Any established tunnel uses 2048 bit ElGamal/Session Tags with 256 bit AES in CBC mode encryption. Signatures use a 1024bit DSA algorithm with a 320 bit seed. TCP connections operate using a 2048 Diffie-Hellman implementation at the moment. Jrandom [5] contains an illustration of the components used in a single tunnel, which happens to be Alice's outbound tunnel and Bob's inbound tunnel for one way communication where Alice is sending a message to Bob.

H. Exit Policy for Internet Communication

I2P can only assure end-to-end privacy and integrity if the two peers involved in communication have joined the I2P network. I2P was never meant to be used for Internet communication. However, some nodes in the network are running exit proxies that enable users to reach destinations located on the Internet.

Similar to Tor, end-to-end encryption with nodes outside the anonymous system are released from the exit nodes without any encryption to reach the destination. Once a reply from the destination sent back to the original sender, the exit node will re-encrypt and repackage garlic cloves to be sent back to the sender. I2P is not meant for anonymous Internet browsing, but for anonymous P2P communication whereby end-users are able to send and receive data anonymously. I2P also supports services similar to Tor's *hidden services* to users in the system. An example of a *hidden service* is websites that are posted anonymously, whereby their IP and ID are not revealed to visitors. Websites are given the extension (**.i2p)*. Although *hidden services* are not purely the focus of this research; however, a few remarks are commented on in the critique section for I2P.

 ISBN:978-1-61208-113-7

H. Threat model

I2P protects against a number of attacks such as: Brute force attacks, Timing attacks, Intersection attacks, Denial of service attacks, Tagging attacks, Partitioning attacks, Predecessor attacks, Harvesting attacks, Sybil attacks, Cryptographic attacks, Development attacks, and Implementation attacks.

I. Critique

I2P is by far the most complicated and most promising anonymous P2P system for many aforementioned reasons. I2P is the product of years of continuous development by a number of dedicated developers that have conducted enough research on existing anonymous systems to come up with a new decentralized system that offers better anonymity to users. Nevertheless, I2P was never meant to be used outside the participating nodes in the system itself. Hence, users connecting to I2P and wishing to browse the Internet or carry out other Internet tasks like chatting, sending emails, and talking using VoIP are not advised to do so. The reason behind this is simply because I2P was never designed for communication between an anonymous system and the Internet. Although there have been some assigned nodes with gateways to the Internet, the nodes are limited in number and are well-known to I2P users.

Just like any anonymous system, I2P has its strengths and weaknesses. The advantages of I2P are:

1) **Message-based instead of Circuit-based:** I2P is a message-based system whereby data packets generated be senders and receivers are encrypted and then wrapped randomly together (using the garlic protocol) and then sent across the I2P network whereby cloves bounce through random hops before reaching their final destination. In comparison with circuit-based, messages no longer need wait for a peer to establish a tunnel through other peers before proceeding with data delivery. Messages simply traverse pre-created tunnels through available nodes on the system. This immensely reduces the overhead of creating tunnels and adds randomness in the system while allowing hops (or router nodes) to control data delivery and add variable latency as well as padding to messages.
2) **Various Protocols Support:** While other anonymous systems restrict Internet communication to a number of protocols like HTTP, and hidden publishing services, I2P offers a wide range of internal services like P2P services (BitTorrent clone), anonymous hidden services like web publishing, anonymous SMTP, file sharing, and anonymous chatting. These protocols have been developed by enthusiasts as plug-ins on top the I2P system that enable a user logged in to I2P to communicate with other users anonymously.
3) **New P2P Infrastructure over the Internet:** The previous point may give an insight on the future of P2P file sharing and communication on the Internet. During these times, the P2P file sharing activities on the Internet have been criticized for housing and transferring illegal and copyrighted content amongst peers in the community. BitTorrent, eMule, Gnutella, and many others are now considered gateways to tremendous amounts of illegal (and legal) content that is being tracked, along with P2P participants, by copyright entities such as the RIAA, DMCA, Interpol, and other similar organizations. Anirban Banerjee, Michali Faloutsos, Laxmi N. Bhuyan's study [15] of P2P file sharing and communication tracking, has shown that file sharing communities now form *block-nets* in order to block certain known legal entities that have been identified using various methods. *Block-nets* are composed of a list of suspected IPs, circulated amongst different P2P systems and regularly updated, for which users trying to communicate from this list of IPs are blocked from participating and joining P2P systems due to their activities in locating and tracking P2P end users. There have been some documented incidents where a number of users have been tracked down and sued over a number of copyright violation using commonly used P2P systems. By introducing I2P as an infrastructure for P2P protocols for anonymous communication, P2P activities would therefore become anonymous to all participating parties. With header and payload encryption, as well as resource hiding, multi-packet routing, and content distribution, entities wishing to track communication will be find it difficult to do so and will be spotted and hence blocked (based on *block-nets*). An example of such a system is *Freenet* [9]. The aim of course is not to allow illegal content to be freely distributed over the Internet; however, to anonymize communication for users on the system. This idea is not bullet proof, as P2P trackers may login to the system anonymously using another set of IPs and conduct various types of attacks, irrelative of time and cost, to pinpoint and breach the system, if need be.
4) **Open Design, Open Source Code:** The developers of I2P, led by *jrandom*, have kept their identities anonymous for many reasons one of which to indicate that the system is anonymous not only in its functionality but in its development also. The system has an open design whereby as enthusiast may login to the I2P website and check the recent and previous developments of I2P and participate in forums while remaining anonymous. The Java

 ISBN:978-1-61208-113-7

source code for the I2P client and server software are freely available for users to inspect and audit if need be. This adds more trust in the system and for adopting the system especially that the system is not sponsored by any government agency as in Tor.

5) **Different Encryption Techniques:** I2P employs a good set of algorithms ranging from symmetric, asymmetric, singing and hashing algorithms that have been used primarily to hide the identity of users in the system and to ensure the integrity of data delivered to peers along the communications paths. With 2048 bit encryptions keys and newly introduced session tags to identify communication amongst peers, as well as defend against replays, I2P stands out to be one of the few anonymous system encompassing such a large number of encryption technologies.
6) **Distributed and Decentralized System:** By making use of a DHT implementation based on Kademlia, as well as removing centralized entities for managing the nodes on the network, the I2P system is protected against attacks on its directory servers. The I2P system is a self-healing and self-organizing anonymous system that is able to keep track of nodes in the system while part of the network is under attack. Moreover, attacking random/specific nodes in the system appears useless as there is no single entity handling I2P system management. Additionally, if *global adversaries* are able to block known I2P nodes for users to connect to, then it will only require another (unknown to *global adversaries*) user node to join the I2P network and then announce, through different means, that it was capable of joining the network and that node itself will become a gateway for new nodes wishing to join the I2P network.
7) **Different Types of Unidirectional Tunnels:** Almost every type of anonymous system that was designed and implemented uses a single tunnel for moving data back and forth from sender to receiver, and hence encrypted tunnels where multi-directional carrying not only encrypted data payloads, but also instructions to other nodes in the system. I2P designers have decided to separate and segregate the role of tunnels by introducing a new set of tunnels such as one for sending data (along with some instructions), one for receiving data (along with some instructions), and another for exploring the network. The latter uses nodes with less bandwidth on less reliable/slower connection nodes. The significance of this segregation is to enhance the amount of peers participating in communication and therefore increase the number of hops along the path of data being transmitted and received. This modification adds a minimum of twice the number of participating nodes in any communication as compared to Tor, which utilizes a single tunnel for sending and receiving data streams as well as instructions to other nodes.
8) **End User Node Participation in Communication:** The new design for I2P encourages that every node joining the I2P needs to use part of its bandwidth as a relay node and pass data to other peers in the system. This approach not only makes use of the end users, who are usually the majority of the nodes in the system using the system, but also adds more hops to any communications that allows more randomness when choosing hops.

The disadvantages of I2P are:

1) **Vulnerability to Partitioning Attacks:** Since I2P utilizes Kademlia for maintaining the distributed system and keeping nodes in contact (using NetDB). Kademlia is susceptible to partitioning attacks that may disconnect targets in the system and therefore reveal the identities of all parties involved in a communication stream. A partitioning attack is an attack that aims at directing end users in the system to connect to a smaller set of malicious nodes only (smaller relative to the complete set of nodes in the system), whereby the malicious nodes are able to simulate the functionality of the anonymous system to the target node, and whereby a user is still able to establish a number of tunnels and select multiple hops. However, all identities for the sender and receiver are actually compromised along the pathways as the nodes participating in communication are malicious nodes. Strong adversaries are also capable of deliberately blocking certain destinations. Hence, other legitimate nodes in the system; and therefore, disconnecting the target at once from the rest of the nodes in the system and then introducing malicious alternative nodes with another set of *NetDB* options and routes. This attack coupled with other types of attacks such as Sybil and Time attacks may fully exploit the identities of the senders and receivers in the system, as well as the data content, especially if one of the malicious nodes is used as an exit node to the Internet.
2) **Possible Intersection Attacks:** An intersection attack is an attack that monitors a certain target and then watches the amount of nodes present and connected to the system constantly. Due to tunnel rotation and variation in target reachability, the attacker may eliminate nodes that have not participated in communication with the target until the target's paths are narrowed down. This will also leave a set of nodes that form these paths

 ISBN:978-1-61208-113-7

exposed for monitoring - seeing as a message traverses the paths from source to destination and vice versa the node is detected. Intersection attacks are strengthened immensely when coupled with other types of attacks such as Sybil and Timing attacks.

3) **Lack of Node and Bandwidth Monitoring:** I2P is a decentralized and distributed system that does not keep track of overall network bandwidth usage and monitoring except for client and router nodes that self-monitor themselves. I2P nodes are capable of storing and graphing their own connection as well as downloading *NetDB* information that are offered by other nodes in the system. One might question the possibility for having an overview of system performance, network bottlenecks on certain nodes, and total bandwidth as well as the number of participating entities in an I2P network (or a graph representation of the connected nodes). Some I2P security enthusiasts argue that revealing such information might risk anonymity as well as reveal weak points in the system. Moreover, as certain peers and routers in the network can variably change their relay capabilities (bandwidth options offered to other peers in the system for relaying their traffic), and as random joins and departures of nodes occur, it might be tough to graph the network or reveal overall relay information without having a centralized polling system to keep track of such frequent change in information. Given that I2P is a decentralized system then the possibility becomes hardly possible, and hence per-node resource and network performance is kept in each I2P node's NetDB information that have decided to participate in the system. On the other hand, when a wide distributed attack is carried out on most of the nodes, and when nodes become unreachable, then would one (or the system) analyze this type of misbehavior and consequently categorize the incident as a false positive, false negative, a single or multiple nodes experiencing connection problems or failure, a DoS attack, a partitioning attack, or any type of network related attack? How would developers in the system realize that router nodes have been flooded with tunnel connections by a large number of peers in a resource consumption attack, whereby fake packet generation (gradually to camouflage the attack and can usually span days and even weeks) can flood the entire network and hence leave the attack to appear as normal network congestion due to a large number of joining parties while in fact it would be a variant of a Sybil Attack?

4) **NetDB Conflicts and Resolution:** The previous point mentions that NetDB stores node information relayed to different clients and routers in the system whereby it contains information about tunnel, node reachability, and encryption information. Now consider the case where a client node connects to the I2P network. During its network investigation for available nodes.

5) DoS Attacks: The I2P team identified three types of attacks that the system can suffer from, and for which the solutions are questionable. The following briefly explains the attacks:

a. Greedy User Attacks: This is actually is not a form of malicious attack on the system, but more associated with a depletion of available relaying bandwidth.
b. Starvation Attacks: The attack is similar to a Sybil Attack whereby nodes joining the I2P network offer connections to other non-malicious nodes in the system. However, after tunnel creation the malicious nodes drop all incoming and outgoing packets to the newly connected nodes. This will cause the nodes to experience frequent network failures as they will not be able to send and receive data using these tunnels. Additionally, this will cause more tunnels to be established with other non-malicious I2P routers due to the lack of connectivity with current malicious nodes.
c. Flooding Attacks: Is an attack that allows a malicious user to introduce a node or a set of node that inject huge amounts of meaningless/meaningful traffic with destinations to inbound and outbound nodes of different peers in the system. Similar to a network DoS attack, nodes in the I2P network receiving such traffic can do nothing to stop this traffic since any node on any network cannot control the amount of traffic it is receiving. I2P developers argue that if nodes detect that huge amounts of traffic are detected, then they can disconnect their tunnels and reestablish new tunnels with other nodes.

IV. CONCLUSION

This paper presented the I2P anonymous-related systems, and their corresponding details that have made them such a success. The paper also commented on the pros and cons of I2P's implementation.

Avoiding traffic analysis and hiding the identities of users is the aim of any anonymous system. However, since most anonymous systems rely on aging encryption technologies for which global adversaries are a capable of compromising, then the integrity of data might be at stake.

This paper also introduced vital topics that need to be further researched such as creating virtual interfaces for

 ISBN:978-1-61208-113-7

making all types of traffic to traverse anonymous systems, as opposed to socket proxies, in order to maximize securing the identity of users on the system and support the widest types of applications possible. Such virtual interfaces can exist in order to ease selecting which types of traffic can pass through the anonymous system and which can be bypassed to leave the anonymous system's utilization at optimum levels.

One of the key elements that worry anonymous systems researchers is QoS for the bandwidth utilized by peers on the systems and the overall network performance. Although this has been slightly commented on, more research in QoS and a bandwidth-choking approach is required while concentrating on security and functionality implications.

In the near future, we plan to focus our work on avoiding traffic analysis and at the same time assuring data integrity using a quorum-based approach. We plan to introduce this work to different anonymous systems researchers and communities for a possible implementation and real testing on existing systems.

ACKNOWLEDGEMENT

This work was funded by the Lebanese American University.

REFERENCES

[1] D. Chaum, "Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms," Communications of the ACM, vol. 24, no. 2, Feb. 1981, pp.84-88.

[2] R. Dingledine, M. J. Freedman, and D. Molnar, "Peer-To-Peer: Harnessing the Power of Disruptive Technologies," Free Haven Project. Retrieved on December 14, 2010 from http://freehaven.net.

[3] O. Sandberg, "Distributed Routing in Small-World Networks," Retrieved on February 10, 2006 from http://torr.eff.net.

[4] The Anonymizer. Frequently Asked Questions, Retrieved June 14, 2010, from http://anonymizer.com.

[5] J. Jrandom, "I2P Anonymous Network: Technical Introduction," Retrieved on December 13, 2010, from http://www.i2p2.de/techintro.html.

[6] J. Jrandom, "Introducing I2P: A Scalable Framework for Anonymous Communication," Retrieved December 13, 2010, from http://www.i2p.com.

[7] R. Dingledine and N. Mathewson, "Tor Path Specification," Retrieved on December 13, 2010 from http://torr.eff.net.

[8] M. Freedman and R. Morris, "Tarzan: A Peer-to-Peer Anonymizing Network Layer," Retrieved on December 14, 2010, from http://www.cs.princeton.edu/~mfreed/docs/tarzan-ccs02.pdf.

[9] I. Clarke, S. Miller, T. Hong, O. Sandberg and B. Wiley, "Protecting Free Expression Online with Freenet," IEEE Internet Computing, Retrieved on December 13, 2010 from http://torr.eff.net.

[10] R. Thommes, and M. Coates, "BitTorrent Fairness: Analysis and Improvements," Retrieved on December 13, 2010, from http://www.tsp.ece.mcgill.ca/Networks/projects/pdf/thommes_WITSP05.pdf.

[11] D. Goldschlag, M. Reed and P. Syverson, "Hiding Routing Information," Retrieved on December 13, 2010, from http://stinet.dtic.mil/cgi-bin/GetTRDoc?AD=ADA465075&Location=U2&doc=GetTRDoc.pdf.

[12] S. Katti, J. Cohen and D. Katabi, "Information Slicing: Anonymity Using Unreliable Overlays," Retrieved on December 13, 2010, from http://nms.lcs.mit.edu/~dina/pub/slicing-nsdi.pdf.

[13] Y. Guan, X. Fu, D. Xuan, P. Shenoy, R. Battati, and W. Zhao, "NetCamo: Camouflaging Network Traffic for QoS-Guaranteed Mission Critical Applications," Retrieved on December 13, 2010, from http://ieeexplore.ieee.org/iel5/3468/20237/00935042.pdf.

[14] P. Maymounkov and D. Mazieres, "Kademlia: A Peer-to-Peer Information System based on XOR Metric," Proceedings of the First International Workshop on Peer-to-Peer Systems - IPTPS. March 2002.

[15] A. Banerjee, M. Faloutsos and N. Laxmi, "P2P: Is Big Brother Watching You?," Retrieved on December 13, 2010 from http://www1.cs.ucr.edu/store/techreports/UCR-CS-2006-06201.pdf.

 ISBN:978-1-61208-113-7

Experimental IPTV and IPv6 Extended Provisioning in a Virtual Testbed

Shuai Qu, Jonas Lindqvist, and Claus Popp Larsen
Netlab
Acreo AB
Hudiksvall, Sweden.
E-mail: {Shuai.Qu, Jonas.Lindqvist, Claus.Popp.Larsen}@acreo.se

***Abstract*—The increasing interest in Internet Protocol Television (IPTV) and Internet Protocol version 6 (IPv6) has driven the need to find a solution to run IPTV and IPv6 outside of a normal public access network. To find new solutions a virtual testbed based on Virtual Private Network (VPN) was built. VPN is a technology that can provide global networking and extended geographic connectivity, also with native security features and good control of the VPN clients. An experimental solution "Virtual Testbed" based on VPN technology to extend provision of IPTV and IPv6 is presented in this paper. This solution allows remote end users over a wider geographical area to participate in IPTV and IPv6 trial since the public network can be used. This also allows researchers an easier access to the test pilots home network and customer premises equipment (CPE), thus makes user behavior research and traffic measurements possible. Evaluation and performance tests on this solution are also illustrated and discussed.**

Keywords - Virtual Testbed, IPTV, IPv6, VPN.

I. INTRODUCTION

IPTV provides digital television services over IP for residential and business users at a lower cost. IPTV is believed to be a killer application for the next-generation Internet and will provide exciting new revenue opportunities for service providers [1]. However, provisioning the IPTV service brings forth significant new challenges [2]. Many commercial IPTV platforms are conventionally deployed on a certain designated network infrastructure with video servers strategically placed. The coverage area of the IPTV network is dedicated. Therefore, it is difficult to distribute IPTV for remote end users who are not part of an IPTV enabled network or in a separate network. It is also costly to operate and extend IPTV at a very big scale. Therefore, a traditional IPTV delivery scheme does not meet these challenges that will be faced in the future, and this drives the need of a solution to extend IPTV provisioning [3].

The amount of home network, as well as the number of services and hosts in them, is increasing. Often the home users cannot get public IPv4 network allocations from service provider and are forced to use Network Address Translation (NAT) to solve connectivity issues to different home services. The need to reach home services from foreign networks has gradually increased as network attached storage, personal video recorders etc. obtain IP connectivity [4]. Therefore, IPv6 is a possible solution to enhance terminal-to-terminal and terminal-to-services connectivity for CPE. It is also beneficial for deep measurement on user behavior and network traffic inside home network in future. Therefore, extended provisioning of IPTV and IPv6 service solution beyond an IPTV-enabled access network, allows for wider access to a "Virtual Testbed" for various test and demonstration purposes. This is extremely useful for our own testbed activities, and we believe it will be of interest to others as well.

Compared to scalabilities, the traditional platform for live experimentation has been physical testbeds: leased lines connecting a limited set of locations [5]. They are more dedicated and are not suitable to extend IPTV provision and IPv6 connectivity for a live experiment.

The Acreo National Testbed (ANT) is built on the fiber infrastructure of the local municipality network "Fiberstaden" in Hudiksvall in Sweden. Commercial and pre-commercial transmission equipment from different vendors has been used to interconnect sites spaced far apart or with high capacity requirements. The broadband access network itself was designed with commercial Layer 2 and 3 equipment (Ethernet switches and Internet routers), also from different vendors [6]. There are around 60 households comprising end-users living in Hudiksvall, and these households are supplied with Internet access and IPTV via Fiber to the Home (FTTH). As a result of geographic limitation, these test pilots are "static" and can only access to network services in ANT locally. It is difficult to extend IPTV and IPv6 services provision to remote users who are not part of ANT network [3]. To address those issues, a *virtual testbed* solution is proposed and has been implemented in a small scale to provide experimental IPTV and IPv6 extended provisioning. More specifically, this paper narrow down to addressing two services issues below:

- Can a solution extend IPTV services provision to users who are not part of a testbed network?
- Can a solution extend IPv6 connectivity to these users as above?

VPN is a generic term that covers the use of public or private networks to create groups of users that are separated from other network users and that may communicate among them as if they were on a private network [7]. VPN can extend geographic connectivity, provide global networking opportunities and reduce operational costs for remote users versus traditional Wide Area Network (WAN). These benefits can facilitate a flexible and cost-effective way to

 ISBN:978-1-61208-113-7

extend IPTV and IPv6 provision. Therefore, IPTV and IPv6 over VPN is an ideal solution to address the issues.

There is also another overlay network technology Peer-to-Peer (P2P), which also can be used for scalable IPTV distribution. P2P does not rely on dedicated network infrastructures and multicast servers. The P2P clients and their connections form an overlay network to exchange video content cooperatively between peers by leveraging their uploading capacity [8]. While P2P network traffic will not fully go into and pass through a specified network, which results in the difficulties in central management, network traffic measurements and some security issues. Comparing to P2P network, VPN network is more centralized and can be configured to make all VPN clients' traffics go through the VPN server. Additionally, VPN uses a flexible user management based on certification system by simple creating or revoking different certificates for different groups of users to achieve users control and authentication. Therefore, VPN solutions are more suitable for our experimental case in this paper than P2P solutions.

Figure 1 illustrates an example of basic IPTV service over VPN. The central office offers IPTV service to different end-users over VPN connections. The IPTV distributions are not constrained by geographic locations.

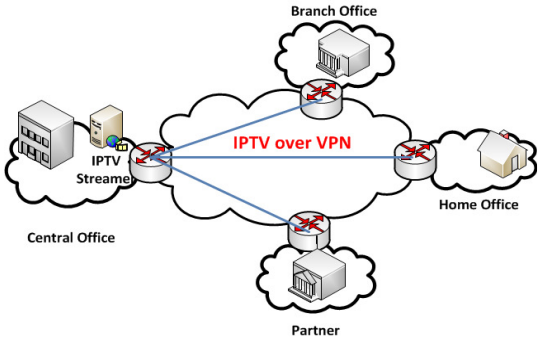


Figure 1. An example of IPTV service over VPN

IPTV over VPN is able to reduce operation costs, transportation costs, provide improved security and better user control [3]. In addition, IPTV over VPN can also provide classified IPTV service features according to geographical groups and customers' demands [9], classified IPTV group services features [9], etc.

One way to provide IPv6 connectivity between end-user sites (when native IPv6 service does not exist) is to use IPv6-over-IPv4 encapsulation (tunneling) between them. The technique encapsulates IPv6 packets within IPv4 so that they can be carried across IPv4 routing infrastructures [10].

Compared to other large virtual testbeds such as PlanetLab [11], our trial is small scale and centralized. The infrastructure is that a VPN network connects one central VPN server, VPN clients and ANT network. Two services, IPTV and IPv6, are running on the same infrastructure. A end user is authenticated with a general client mechanism to be a legal test pilot. Our trial is user-orientated, which provides services directly to end users. This also allows researchers an easier access to the test pilots home network and CPE, thus makes user behavior research and traffic measurements possible.

The contributions in this paper are threefold: 1) One virtual testbed solution is proposed to extend a trial IPTV and IPv6 provision. In principle, people all over world who have broadband connections can access to this virtual testbed and participate in our IPTV and IPv6 services trial over VPN tunnels. 2) The traffic measurements have been performed and the results showed that a VPN solution can provide IPTV and IPv6 with acceptable Quality of Service (QoS) to remote end users. 3) All implementations are built on different kinds of open source software, which makes these services more economical and cost-effective. The rest of this paper is organized as the follows. The proposed scheme is presented in Section 2. Section 3 describes how experiments are designed to implement proposed scheme. Section 4 presents the evaluations and test results. Conclusion is made in Section 5.

II. PROPOSED SCHEME

There are different kinds of VPN technologies, e.g., Point-to-Point Tunnel Protocol (PPTP) VPN, Layer 2 Tunnel Protocol (L2TP) VPN, IPSec VPN and OpenVPN. Those different VPN technologies will be analyzed to find a suitable solution to deliver IPTV and IPv6 over VPN for our case.

A. *State of the Art*

Some standards and specifications about how to deliver IPTV and IPv6 over VPN have been designed. In "MPLS and VPN Architectures Volume II" [12], Chapter 7 "Multicast VPN", defines multicast VPN and introduces some Multicast VPN examples. "Multicast over IPsec VPN Design Guide" [13] was published by Cisco System gives a detailed guide to implement Multicast distribution over IPsec VPN network based on Cisco switches and routers. The Internet Draft "Multicast in MPLS/BGP IP VPNs" [14] was written by engineers at Cisco and describes the MVPN (Multicast in Border Gateway Protocol (BGP)/Multi-Protocol Label Switch (MPLS) IP VPNs) solution with Cisco equipment. "ITU-T IPTV Focus Group Proceedings" [9] promotes the global IPTV standards. In other aspect part of the standards, the Work Group (WG) 3 has identified some requirements on Multicast VPN in IPTV network Control and Multicast VPN Group Management aspect. The Internet Standards Track "Transition Mechanisms for IPv6 Hosts and Routers" [15] describes the "IPv6-over-IPv4 tunneling" technologies and transition mechanism for IPv6. For these IPTV VPN solutions, some standards focus on MPLS VPNs, which needs at least backbone networks to support MPLS. Some solutions use IPSec VPN, which requires specific vendor's hardware or software, for example, Cisco System to deploy. So those solutions are not so flexible and open to set up for our experiment.

B. *Possible VPN solutions for IPTV VPN*

To find out the most suitable VPN solution, comparisons between different VPN technologies are made as shown below.

 ISBN:978-1-61208-113-7

- From security perspectives, 1) PPTP VPN is vulnerable to man-in-the middle attack and weak in authentication. 2) For lack of confidentialities, L2TP is often implemented with IPSec for data confidentiality. 3) OpenVPN offers the same security functions and features as IPSec does. The IPSec protocol is implemented as a modification of the IP stack in the kernel stack. But the kernel interactions add security risks on the Operation System (OS) [16].
- For packet overhead, IPSec adds an extra size byte to the original packet, which needs more overhead compare to OpenVPN [17].
- For easy usage, the kernel-space based IPSec requires independent implementation for every OS. The user-space based OpenVPN is much easier to be ported to other OS
- For NAT traversal compatibility, OpenVPN only uses one single port for communication, which is extremely firewall-friendly. The Authentication Header's (AH) source address checking mechanism makes IPSec incompatible with the NAT traversal.
- From multicast support perspective, OpenVPN can natively support multicast while IPSec needs to combine the Generic Record Encapsulation (GRE) tunnel to support multicast. The IPSec Direct Encapsulation only supports unicast IP. IP multicast (IPmc) is not supported with IPSec Direct Encapsulation. IPSec was created to be a security protocol between two and only two devices, so a service such as multicast is problematic. An IPSec peer encrypts a packet so that only one other IPSec peer can successfully perform the de-encryption. IPmc is not compatible with this mode of operation [13].
- In addition, IPSec services usually require third-party hardware or software while OpenVPN is open source software, which makes it very cost-effective.

OpenVPN is an open source and user space tunneling package. OpenVPN uses the OpenSSL library to provide encryption of both the data and control channels [18]. Additional benefits of using OpenVPN are:

- tunnel any IP sub-network or virtual Ethernet adapter over a single User Datagram Protocol (UDP) or Transmission Control Protocol (TCP) port [19],
- multiple load-balanced VPN servers farm, which can handle thousands of dynamic VPN connections,
- use security features of the OpenSSL library to protect network traffic,
- use real-time adaptive link compression and traffic-shaping to manage link bandwidth utilization [19]

One problem of OpenVPN is that OpenVPN is mostly a software-only product until now and it is not found in any hardware applications. Although IPSec is supported by most vendors and can be found in the hardware applications (Routers, Firewalls, etc), incompatibilities between different vendors make IPSec painfully difficult to setup. Another problem of OpenVPN is that it is a user-space program using OpenSSL crypto library. OpenVPN handles data packets based on the TCP/UDP tunnel and TUN/TAP virtual network interface. Therefore, OpenVPN is heavier than IPSec in terms of performance. In summary, OpenVPN is a more suitable VPN solution to deploy multicast service over VPN than the others.

C. *IPv6-over-IPv4 tunnelings*

The tunneling concept is to encapsulate an IPv6 packet as the payload of an IPv4 packet [20]. The IPv4 Protocol field is set to type 41 to indicate an encapsulated IPv6 packet. An IPv4 header with Source and Destination IPv4 addresses is added in front of IPv6 packets. The Source and Destination addresses are set to the tunnel endpoints IPv4 addresses. The tunnel endpoints can be manually configured or automatically derived from the sending tunnel interface and the next-hop address of the matching routing for the destination IPv6 address in tunneled packet, so that IPv6 packets can be sent over the IPv4 infrastructure. Figure 2 illustrates that an example of encapsulating IPv6 in IPv4 packet.

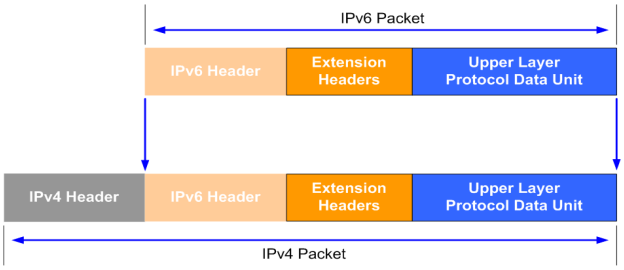


Figure 2. An example of encapsulating IPv6 in IPv4 packet

The IPv6-over-IPv4 tunneling technology also applies to OpenVPN tunnels. Point-to-point IPv6 tunnels are supported on Operating Systems, which have IPv6 TUN driver support (this includes Linux and the BSDs). IPv6 over TAP is always supported as is any other protocol, which can run over Ethernet [21].

III. EXPERIMENTAL SETUP

The experimental implementation is based on ANT infrastructure, which is described in Part I Introduction. The experimental IPTV and IPv6 virtual testbed was built on ANT network infrastructure and designed as shown in Figure 3.

A. *IPTV and IPv6 VPN testbed layout description*

The following descriptions are all related to Figure 3.

- Number 1, IPTV system Hudiksvall.
- Number 2, Acreo Hudiksvall Router: this router is the core router of the ANT project in Hudiksvall.
- Number 3, VPN Server: the VPN Server is linked up together over a VPN tunnel with the VPN individual clients or home gateway. Different open source software was installed on this server. Together with the core router, the VPN server provides IPv4, IPv6, VPN and multicast services to VPN clients.

 ISBN:978-1-61208-113-7

- Number 4, The Public Network.
- Number 5, Home Gateway: the home gateway is physical placed between the VPN server and home network and running on an open source routing platform – OpenWRT [22]. The gateway plays four roles: 1.A VPN client to establish VPN connections. 2. An Internet Group Management Protocol (IGMP) proxy [23] to provide multicast routing. 3. A Router Advertisement Daemon (radvd) [24] to provide IPv6 stateless auto configuration. 4. A home gateway to provide home networking.
- Number 6, Individual VPN clients: the laptops installed the VPN client program.
- Number 7, Different clients inside home network
- Number 8, The IPv6 network
- Number 9, The KAME project (www.kame.net) Server: an IPv6 project named KAME and the server can provide an IPv6 connectivity test.

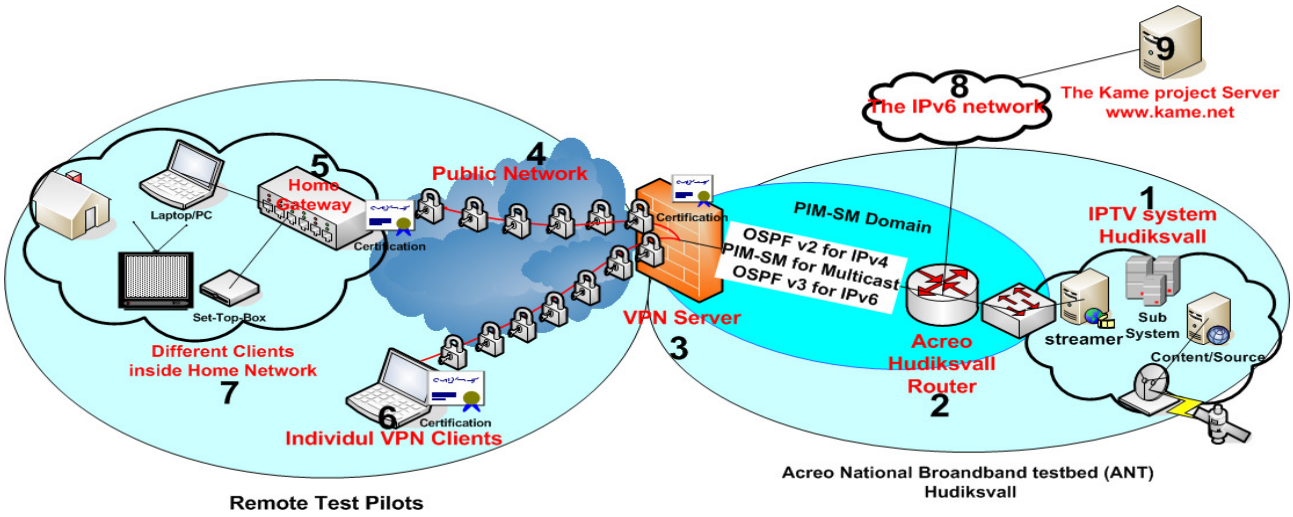


Figure 3. IPTV and IPv6 VPN testbed layout.

There are mainly two implementations: IPTV VPN and IPv6 VPN.

For IPTV VPN, OpenVPN was set up to provide VPN services; Open Shortest Path First version 2 (OSPFv2) was implemented to provide unicast routing; Protocol Independent Multicast - Sparse Mode (PIM-SM) was built up to provide multicast routing; Home gateway was developed to support gateway-to-gateway VPN connections. The home gateway was built on an embedded Linux box with different open source software installed, to establish an automatic VPN connection to the VPN server. In that way IPTV and Internet connections can be provided to the end users in home networks.

The IPTV VPN starts up as follows. For host-to-gateway connections, a laptop with a VPN client program configured connects to Acreo's own VPN server. The server will then set up an IPv4 VPN-tunnel between the server and client. The laptop will then obtain a public VPN IP address via the Dynamic Host Configuration Protocol (DHCP) service, which the VPN server provides. The OSPF v2 and PIM-SM routing protocol are running between the VPN server and Acreo Hudiksvall Router. The internet traffic is routed over the tunnel via the VPN server to the Acreo Hudiksvall Router. The multicast traffic from the source in the IPTV system Hudiksvall is routed via the Acreo Hudiksvall Router (the Rendezvous Point (RP) in the PIM-SM domain) to the VPN server (PIM-SM enabled) over a VPN-tunnel to the client. The differences between the gateway-to-gateway and host-to-gateway VPN connection is the home gateway play three roles of VPN client, IGMP proxy and normal home gateway.

For IPv6 VPN implementations, IPv6-over-IPv4 tunneling and OSPFv3 were implemented to provide IPv6 connectivity and unicast routing. The IPv6 VPN start-up procedure is as follows. For host-to-gateway connections, an end-user's laptop starts up the VPN client program, which automatically establishes an IPv4 based VPN tunnel to the server. The client is manually configured an IPv6 address with the gateway pointed to the VPN server to establish an IPv6-over-IPv4 VPN between the client and server. The IPv6 traffic will then be routed via the OSPFv3 protocol running between VPN server and Acreo Hudiksvall Router to the IPv6 network.

For gateway-to-gateway connection, besides setting up an IPv6 connection from home gateway to the VPN server, the home gateway also provides IPv6 home networking by using radvd. This daemon can listen to router solicitations (RS) and answer with router advertisement (RA). These RAs contains information, which is used for hosts to configure their interfaces and includes IPv6 address prefixes, the link Maximum Transmission Unit (MTU) and default routers information. With the help of radvd, a PC or laptop with IPv6 stack installed is able to automatically configure its interface to appropriate IPv6 address. Any global IPv6 address can be pingable from this home gateway and any IPv6-enabled host in home network. In our trial, only IPv6

 ISBN:978-1-61208-113-7

unicast is deployed. IPv6 multicast is out of scope of this experiment concerns and will not be discussed here.

IV. MEASUREMENTS AND ANALYSIS

Some measurement instruments and methods were used to evaluate the QoS of IPTV and IPv6 over VPN connections. IPTV testing was conducted by one professional IPTV measurement system - Agama Analyzer [25]. VPN connection qualities were measured by two websites, which are world widely-used for broadband speed and quality test. In summary, three test activities shown as below were conducted.

- Evaluate VPN services qualities.
- Compare IPTV over VPN and normal IPTV service qualities.
- A simple test on IPv6 connectivity.

A. *VPN service qualities measurements*

OpenVPN utilizes different cryptographic algorithms to achieve user authentication, authorization, network data confidentiality and integrity. Therefore, as the carrier tunnels for IPTV and IPv6, the QoS impact to the services due to VPN encryption need to be quantified due to its importance to the service quality for distributed users. Fortunately, OpenVPN is flexible to be configured to enable and disable these security options for measurements.

The experiments were conducted on one OpenVPN server and one OpenVPN client as follows, interconnected with high-speed backbone network with capacity above 1 Gbps spanning about 300km.

OpenVPN Server:

1. SUSE Linux Enterprise Server 10 SP2 (x86_64) (Kernel 2.6.16.60-0.21)
2. 2*Intel Xeon(TM) 3.00GHz 64bit CPU, 1GB RAM
3. 2*NIC 10/100/1000M bps, 100 Mbps Switching
4. OpenVPN 2.1_rc18

OpenVPN Client:

1. Windows 7 Professional, 2*Intel Core(TM)2 Duo CPU P8600 2,4 GHz, 4G RAM.
2. NIC 10/100/1000M bps, 100 Mbps Switching
3. OpenVPN 2.1_rc22 and OpenVPN GUI 1.03
4. The Global Broadband Speed and Quality Test websites: speedtest.net and pingtest.net.

Mission-critical IPTV service quality requires sufficient network bandwidth to assure delivery without loss, low network delay and so on. So the following parameters will be measured: network bandwidth, network delay and network capacity loss. The general testing scenario for a VPN client is as following. A Laptop/PC installed an OpenVPN client remotely connect a VPN Server. After establishing VPN tunnel, the client uses speedtest.net and pingtest.net to measure network quality. The OpenVPN server runs in two modes—either over UDP or TCP. The UDP mode is preferred due to better performance, as the UDP mode is not limited by the TCP congestion control algorithm [26][27]. In particular, UDP based VPN for real-time multicast communication shows minimum impact on traffic and slight CPU requirement increase comparing to TCP based VPN mode [27]. Therefore, UDP based VPN mode, with six different options combing a number of network QoS critical parameters, was chosen to conduct the measurements. Since OpenVPN requires extra management, which could lead to a capacity reduction, so option 1 in Table I is defined as original network connection case to measure and compare network capacity loss. In addition, encryption will increase OpenVPN traffic overhead and compression will influence data transmission efficiency [17]. Therefore, from option 2 to option 5 is different combination of those options to measure and verify which option wins the best QoS.

The test case 1 and test case 2 were performed ten times. Two example results checking against speedtest.net and pingtest.net separately are shown in Figure 4 and Figure 5. The two test cases measurement values are presented in Table I as below.

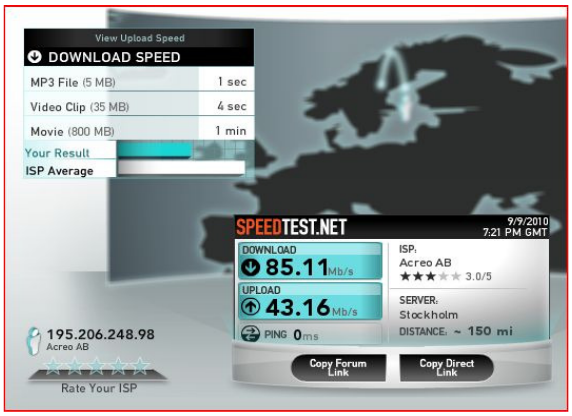


Figure 4. The test case 1, network bandwidth check against speedtestb.net

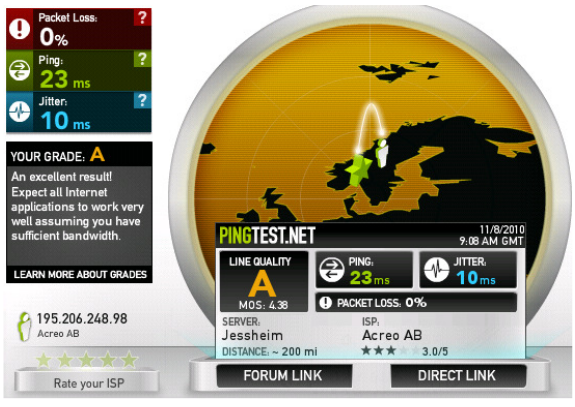


Figure 5. The test case 2, network delay check against pingtest.net

The VPN service connectivity benchmark results can be summarized as follows. 1) The VPN network requires extra management overhead, which leads to a bandwidth reduction. In Table I, comparing to option 1 results, network bandwidth under other options shows that the VPN network bandwidth loss rate is approximate 26%--32%. 2) For network delay, the data compression "comp-lzo" option can

 ISBN:978-1-61208-113-7

reduce VPN network delay while the security options worsen the network delay. In Table I, the VPN connection without security options but with data compression enabled is the winner in all tests. The VPN connection with all security options shows rather larger network delay (average 38.35ms). The mission-critical IPTV service requires low network delay and high real-time multicast traffics. However, encryption of multicast streaming will consume system resource and give negative impact on the service QoS. If there is no confidentiality requirement for multicast streaming, then authentication of both communication parties, to some extent, is able to ensure IPTV security. The consumption of system resource is accordingly reduced and the services performance could be improved.

TABLE I: BANDWIDTH CONNECTIVITY TEST RESULTS FOR UDP-BASED VPN MODE WITH DIFFERENT VPN SERVER OPTIONS. HMAC STATNDS FOR "HASH MESSAGE AUTHENTICATION CODE"

	Option 1	Option 2	Option 3	Option 4	Option 5	Option6
VPN connections		x	x	x	x	x
Encryption		x	x			
Integrity check with HMAC		x		x		
Data compression		x	x	x	x	
Average Download Speed	84.15Mb/s	59.48Mb/s	59.72Mb/s	62.09Mb/s	59.78Mb/s	61.45Mb/s
Average Upload Speed	42.56Mb/s	30.83Mb/s	31.94Mb/s	33.04Mb/s	31.98Mb/s	33.45Mb/s
Average Network Delay	13.01ms	38.35ms	20.47ms	24.44ms	16.23ms	17.36ms
Maximum Download Speed	88.75Mb/s	61.84Mb/s	62.32Mb/s	65.38Mb/s	66.82Mb/s	67.18Mb/s
Maximum Upload Speed	47.72 Mb/s	31.45 Mb/s	32.84 Mb/s	35.13Mb/s	35.44 Mb/s	36 Mb/s
Shortest Network Delay	8ms	8.9ms	8.7ms	8.65ms	8.77ms	8.56ms

B. IPTV VPN service qualities

IPTV service qualities comparisons between IPTV VPN and normal IPTV had been done with this Agama instrument – Agama Analyzer. The measurement is based on the network setting up shown in Figure 6. The client is a laptop without OpenVPN client installed. The switch is 10/100M Fast Ethernet Switch. Agama Analyzer is connected to switch with two clients together to measure IPTV QoS. The further descriptions of other components and service scenarios in Figure 6 can be referred to Part III Experiment Setup Section A. During the experiment, the following parameters were used to qualify the QoS provided [28].

- Packet loss is measured as the portion of packets transmitted but not received in the destination compared to the total number or packets transmitted.
- Packet Jitter is often used as a measure of the variability over time of the packet latency across a network [29]. A bigger number of packet jitter value means larger packet latency.

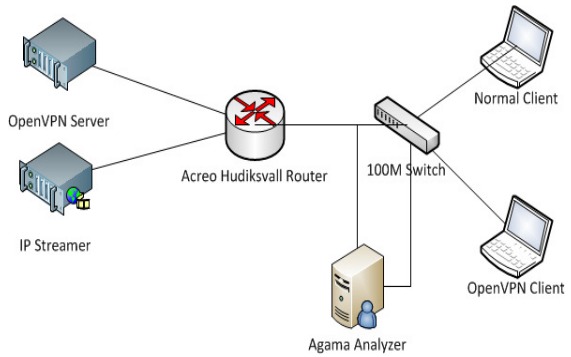


Figure 6. Testbed for IPTV VPN and normal IPTV comparisons

Figure 7, Figure 8 and Figure 9 show that the test results that one IPTV channel with Bitrates around 4Mbit/s from the same streamer was measured by Agama Analyzer during 72 hours (from 2009-05-29 8:00 to 2009-06-01 8:00). The results are summarized as follows: 1.Connection without VPN network bandwidth of 64Mbit/s downstream and 38Mbit/s upstream were achieved. VPN connection network bandwidth of 37Mbit/s downstream and 16Mbit/s upstream were achieved. 2. Over a three days period no noticeable signs of distortion were visually observed by human monitoring.

The Packet Jitter measurement results from Figure 8 and Figure 9 are summarized and presented in Table II as below:

TABLE II: THE PACKET JITTER MEASUREMENT RESULTS FROM AGAMA ANALYZER

	normal wired lines	VPN
Average Packet Jitter	6.1ms	9.6ms
Maximum Packet Jitter	10.1ms	42.3ms

Although there are many suggested video quality metrics with varying degree of performance, most of the more well known e.g., Video Quality Metric (VQM) [30] require access to the original undistorted reference i.e., full reference or reduced reference metrics, which are not in general available. The methods that do not require such access i.e., no-reference metrics have not performed good enough to be standardized. Still we wanted to get on estimate of the performance of the transmission of the IPTV over the VPN connection. The Agama Analyzer analyzes the video stream for consistency and completeness according the codec standard. From this inferences about the likely quality of the transmitted IPTV could be made. They are classified in three levels by the Agama Analyzer. No errors found, which means that the video have the same quality as when it was transmitted. Minor errors found, which will have just minor impact on the quality and then major errors found, which also will have substantial impact on the quality. According to the Agama measurement results, the IPTV has only suffered minor distortions over VPN connections and has only degraded slightly.

 ISBN:978-1-61208-113-7

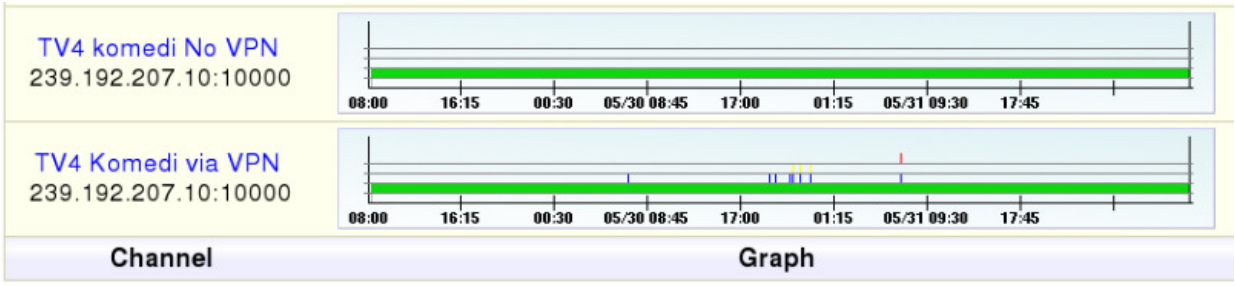


Figure 7. SVT TV4 Komedi channel measuring graph by Agama Analyzer from 2009-05-29 8:00 to 2009-06-01 8:00. Green=OK, Blue=minor distortion, Yellow=major distortion, Red=Packet loss. During the same time, the TV is delivered over the connections without VPN and with VPN connection separately.

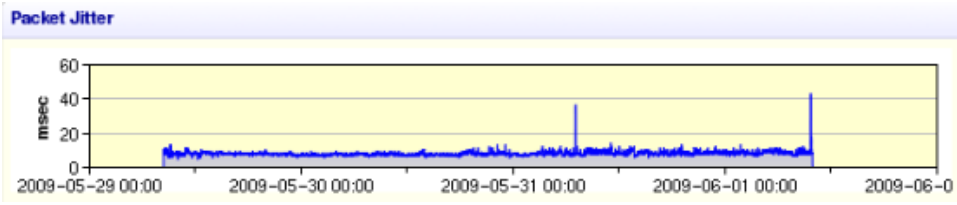


Figure 8. SVT TV4 Komedi channel Packet Jitter measurement results with no VPN connections from Agama Analyzer.

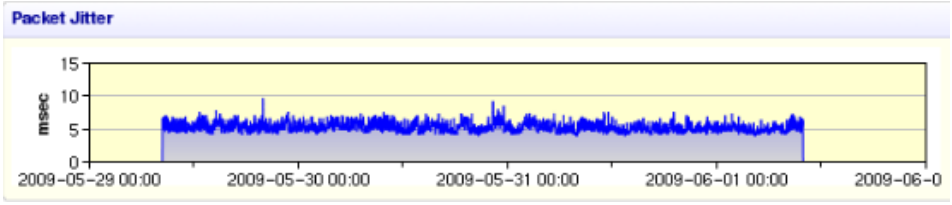


Figure 9. SVT TV4 Komedi channel Packet Jitter measurement results with VPN connection from Agama Analyzer.

C. IPv6 connectivity measurements

Based on IPv6 VPN experiment setup described in Part III, with the help of KAME project server with IPv6 enabled, a simple IPv6 trace route test was performed. A VPN client manually configured with an IPv6 address trace route against www.kame.net. In this paper only network connectivity is concerned, and the IPv6 network quality in terms of packet loss, network delay is not considered here. The result is shown in Figure 10.

The result shows that IPv6 packets were successfully routed from a VPN client over an IPv6-over-IPv4 tunnel to the IPv6 network, and the packets traversed the IPv6 network via the IPv6 enabled network nodes hop by hop to the destination.

V. CONCLUSION AND FUTURE RESEARCH

In this paper, a VPN solution is designed and implemented to realize a "Virtual Testbed", which can extend trial provisioning of IPTV and IPv6 spanning a wider geographical area to remote end users at a lower network operation cost. This is very useful for our testbed activities, which have so far been confined to reach test pilots within the municipality of Hudiksvall open city network. The proposed schema uses proven and standardized technologies, with open solutions and open-source software. This solution makes it very cost-effective and commercially applicable, for potential providers who wants to extend their services. The results of the evaluation showed acceptable service QoS. However, it should be aware that this VPN solution is not always the best solution due to an approximate 30% network capacity reduction. But VPN is still a good way or in some case the only solution to extend IPTV service provision with a centralized network infrastructure. Meanwhile, IPv6 connectivity can also be extended over this VPN infrastructure to achieve good terminal-to-terminal and terminal-to-services connectivity. So with the help of this solution, a virtual testbed can have more scalable access from dynamic test pilots and provide an attractive and extendable platform for IPTV and IPv6 services provision and experimentation.

As part of future work, we intend to conduct pressure and load testing to improve and further optimize performance of our system. We plan to more thoroughly investigate the VPN processing performance with multiple streams over many geographically distributed clients. In addition, our future research also includes VPN connection improvement on a

 ISBN:978-1-61208-113-7

high performance. We expect such extended research work will be able to make our solutions even better.

ACKNOWLEDGMENT

The authors acknowledge the EU Regional development Funds for supporting this work through the project "Acreo National Testbed, phase 2" (ANT2)

REFERENCES

[1] Y. Xiao, X. Du, J. Zhang, and F. Hu, "Internet Protocol Television (IPTV):The Killer Application for the Next-Generation Internet," IEEE Communications Magazine, vol. 45, no. 11, pp. 126 - 134, Nov, 2007.

[2] R. Jain, "I want my IPTV," IEEE Multimedia, vol. 12, no. 3, pp. 96, 2005.

[3] S. Qu and J. Lindqvist, "Scalable IPTV Delivery to Home via VPN," Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Volume 40, Part 10, pp. 237-246, 2010.

[4] K. Huhtanen, B. Silverajan, and J. Harju, "Utilising IPv6 over VPN to Enhance Home Service Connectivity," Terena 2007 Special Issue of the Journal of Campus-Wide Information Systems Vol 24 No 4 pp. 271-279.

[5] L. Peterson, S. Shenker, and J. Turner In Third Works, "Overcoming the Internet Impasse through Virtualization," in Third Workshop on Hot Topics in Networks (HotNets-III), Nov. 2004.

[6] C. P. Larsen, R. Flodin, C. Lindqvist, R. Lindstrm, H. Pathirana, and A. Gavler, "Experience with IPTV in the Testbed," Acreo Broadband Communication Project report Y2002-Y2006: Dec. letter 2004-01780, Acreo AB, pp. 18 – 27, January 31st 2007.

[7] L. Andersson and T. Madsen, "Provider Provisioned Virtual Private Network (VPN) Terminology," Internet Request for Comments RFC 4026, March 2005.

[8] X. J. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross, "A Measurement Study of a Large-Scale P2P IPTV System," IEEE Transactions on Multimedia, vol. 9, no. 8, pp. 1672 – 1687, Dec. 2007

[9] ITU-T, "ITU-T IPTV Focus Group Proceedings," pp. 389 – 390, 2008.

[10] B. Carpenter, B. Fink, and K. Moore, "Connecting IPv6 Routing Domains Over the IPv4 Internet," The Internet Protocol Journal, vol. 3, no. 1, pp.2-10, March 2000.

[11] PlannetLab, http://www.planet-lab.org/, 10.11.2010.

[12] I. Pepelnjak, J. Guichard, and J. Apcar, "MPLS and VPN Architectures Volume II," *Cisco Press*, pp. 333 – 387, 2003.

[13] "Multicast over IPsec VPN Design Guide," www.cisco.com/en/US/docs/solutions/Enterprise/WAN_and_MAN/V3PNIPmc.html , 11.11.2010.

[14] E. Rosen, Y. Cai, and J. Wijsnands, "Multicast in MPLS/BGP VPNs," Internet Draft, August 18, 2009.

[15] R. Gilligan and E. Nordmark, "Transition Mechanisms for IPv6 Hosts and Routers (RFC 2893)," IETF, August, 2000.

[16] C. Hosner, "OpenVPN and the SSL VPN Revolution," SANS Institute, pp.10, Aug 2004.

[17] A. Alshamsi and T. Saito, "A Technical Comparison of IPSec and SSL," 19th Intl. Conf. on Advanced Information Networking and Applications (AINA'05), vol. 2, pp. 392–395, Mar. 2005.

[18] D. H. Ryu and S. H. Nam, "Implementation of wireless VoIP system based on VPN," Proc. The 7th WSEAS Int. Conf. on Electronics, Hardware, Wireless and Optical Communications, Cambrige, UK, Feb. 2008.

[19] "What is OpenVPN," http://www.openvpn.net/index.php/open-source/333-what-is-openvpn.html, 11.11.2010.

[20] D. C. Lee, D. L. Lough, S. F. Midkiff, N. J. Davis and P. E. Benchoff, "The Next Generation of the Internet: Aspects of the Internet Protocol Version 6," IEEE Network, January/February 1998, pp. 28 - 33.

[21] "Is IPv6 support planned/in the works?" http://openvpn.net/index.php/open-source/faq/77-server/287-is-ipv6-support-plannedin-the-works.html, 10.11.2010.

[22] OpenWrt – Wireless Freedom, http://www.openwrt.org, 11.11.2010.

[23] C. Cho, I. Han, Y. Jun and H. Lee, "Improvement of Channel Zapping Time in IPTV Services Using the Adjacent Groups Join-Leave Method," Advanced Communication Technology, the 6th International Conference on, Vol. 2, pp. 971 – 975, 2004.

[24] P. Vidales, G. Mapp, F. Stajano, J. Crowcroft, C.J. Bernardos, "A Practical Approach for 4G Systems: Deployment of Overlay Networks," *TRIDENTCOM'05*, pp. 172 – 181, 2005.

[25] Agama Analyzer, Agama Technologies AB, Box 602, SE-581 07 Linkoping, Sweden, 2010.

[26] V. Jacobso, "Congestion avoidance and control," ACM SIGCOMM '88, Stanford, CA (1988) 314–329

[27] P. Holub, E. Hladka, M. Prochazka, M. Liska, "Secure and Pervasive Collaborative Platform for Medical Applications," IOS PRESS, Studies In Health Technology and Informatics, 2007, VOL 126, pp. 229-238.

[28] IP Performance Metrics (IPPM) Working Group, IETF, http://www.ietf.org/html.charters/ippm-charter.html, 11.11.2010.

[29] D. H. Wolaver, "Phase-Locked Loop Circuit Design," Prentice-Hall, ISBN 0-13-662743-9, pp. 211-237.

[30] M. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality", IEEE Transactions on Broadcasting , vol. 50, Issue. 3, pp. 312-322 , Sept. 2004.

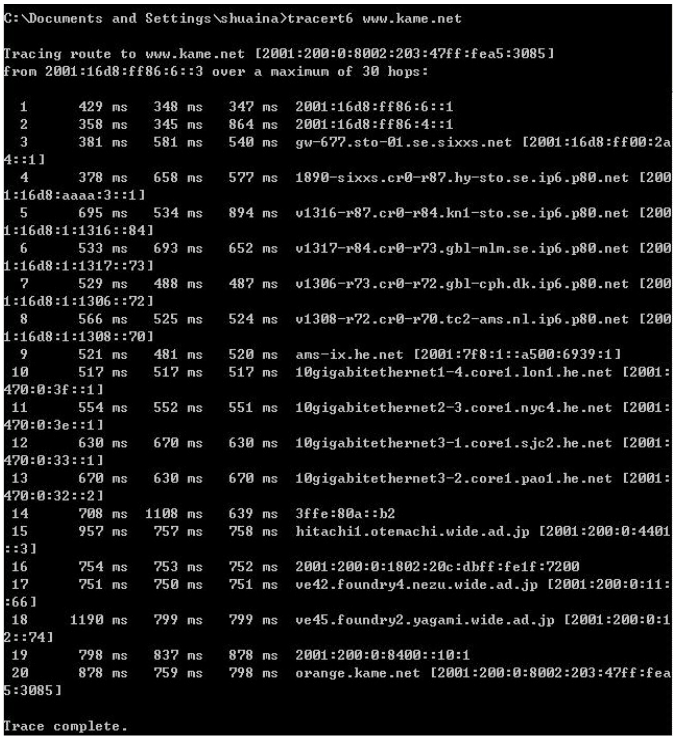


Figure 10. The IPv6 trace route check against www.kame.net

Analysis of the Implementation of Utility Functions to Define an Optimal Partition of a Multicast Group

Joël Penhoat, Karine Guillouard
France Télécom
Issy-les-Moulineaux, France
{joel.penhoat; karine.guillouard}@orange-ftgroup.com

Tayeb Lemlouma, Mikaël Salaun
Université de Rennes 1
Rennes, France
{tayeb.lemlouma; mikael.salaun}@univ-rennes1.fr

***Abstract*— As the optimal use of network resources is a major issue for telecoms operators, we started works aiming to, firstly, improve the utilization of network resources by transmitting the IP packets in multicast when possible, secondly, to adapt the format of the data transmitted in multicast to take into account the context of the members of the multicast group, and thirdly to preserve the Quality of Service when a member of the multicast group moves from a radio network to another radio network. The paper shows, through a scenario, how our work will improve the utilization of the resources and then describes our approach.**

Keywords- Handover; Multicast group partition; Resource optimization; User's context; Utility function

I. INTRODUCTION

The optimization of the use of the resources of the networks is a major issue for a telecoms operator because it allows him to reduce his operational expenditure (OPEX). In particular, the improvement of the use of the resources of the radio spectrum is necessary as showed in a study [1] led in 2002 by Federal Communications Commission (FCC). This study showed that in a frequency band the rate of use of the radio resources could vary between 5% and 85%.

According to us, the solutions that aim to optimize the use of the network resources must be implemented in each level of the TCP/IP stack. Moreover, interactions must exist among each solution implemented in each level of the stack, i.e. the optimization is based on cross-layer solutions. In the PHY and MAC layers, the Cognitive Radios [2] are designed to improve the use of the radio spectrum resources by exploiting the radio resources vacated by their owners. In 2004, the FCC has asked the IEEE to implement the Cognitive Radios in the frequency range 54 -698 MHz [3]. The IEEE P802.22 standard [4] meets this demand by allowing the use of vacant TV channels by radio equipments operating without radio licenses. In the IP layer, the multicast transmission of data [5] is a known technique for improving the use of the resources in an IP network because it reduces the number of IP packets transmitted over a network when several receivers must receive the same data. In the transport layer, numerous studies have improved the throughput of TCP by taking into consideration the physical characteristics of the networks. For example, the TCP Westwood protocol [6] improves the throughput of a TCP connection when the IP packets are transmitted over a radio network.

Our study focuses on the improvement of the use of the resources of networks at the IP level by implementing multicast transmissions when the services asked by the users can be transmitted in multicast. During a multicast transmission, all the members of a multicast group (i.e. the users) receive the data in the same format because a multicast transmission does not take into account the heterogeneity of the receivers (mobile phones, laptops, smartphones, tablets), the heterogeneity of the radio networks (GSM, Wi-Fi, WiMAX ...), or the diversity of the profiles of the members of the group (engineer, accountant, student …). The diversity of the receivers, radio networks, or the profiles of the users, are elements that characterize the context of a user. In the article, we use the definition of Abowd *et al* [7] to define the context of a user. Figure 1 illustrates a scenario in which the members of a multicast group G have different contexts. In addition, as the members of the group may be mobile, their contexts vary during the transmission because they can connect themselves to wireless networks having different characteristics, for example during a handover between a Wi-Fi network transmitting in multicast the IP packets and an Universal Mobile Telecommunications System (UMTS) network transmitting in unicast the IP packets.

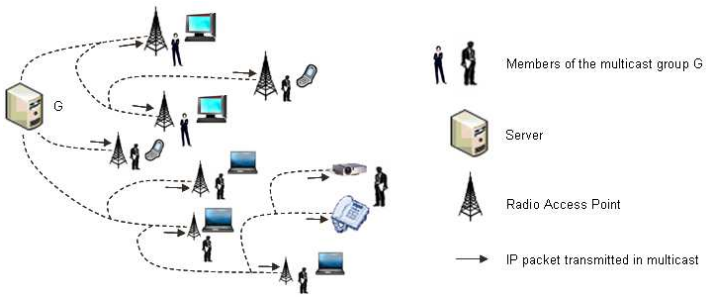


Figure 1. Multicast transmission in a heterogeneous context

So, the three objectives of our work are, firstly, to improve the utilization of the resources of the networks by transmitting in multicast the IP packets when possible, secondly, to adapt the format of the data transmitted in multicast to take into account the context of the members of the multicast group, and thirdly to preserve the Quality of

 ISBN:978-1-61208-113-7

Service (QoS) when a user moves from one radio network to another radio network.

The structure of the article is the following. The second section describes a scenario showing the interest of our works for the telecoms operators. The third section makes an inventory of the works that take into account the context of the users during a multicast transmission and analyzes their loopholes; then the fourth section presents our approach to implement our three objectives. Finally, the fifth section concludes the article and exposes our future works.

II. DESCRIPTION OF A SCENARIO ESTABLISHING A SYNERGY BETWEEN UNICAST AND MULTICAST NETWORKS

In this section we describe a scenario that could improve the utilization of the resources of the radio networks of a telecoms operator by establishing a synergy between unicast and multicast networks. In our scenario, eight persons, Anatole, Antoine, Bernard, Bertrand, Alice, Bénédicte, Catherine, and Isabelle, take part in a video-conference. The video-conference is made up of a Voice over IP (VoIP) flow and a Video flow. The Video flow is encoded with a Scalable Video Codec [8] that splits the flow into several sub-flows. Our scenario consists of six stages. In the first stage (Figure 2), Anatole and Antoine receive the VoIP flow on a fixed phone and the Video flow on a video-projector.

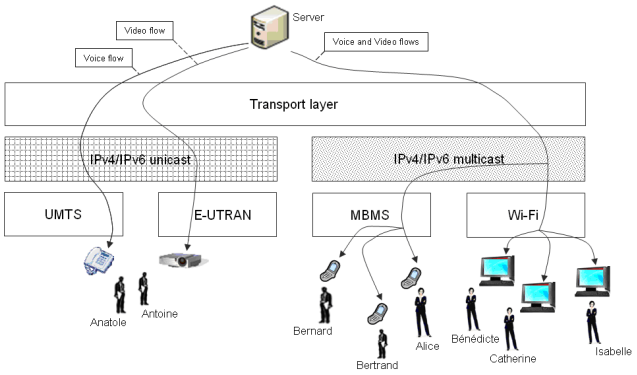


Figure 2. A scenario describing a synergy between multicast and unicast networks (first stage)

The VoIP flow is transmitted in unicast over a UMTS network that uses IPv4 addresses. The Video flow is transmitted in unicast over an Evolved Universal Terrestrial Radio Access Network [9] that uses IPv6 addresses. The Evolved Universal Terrestrial Radio Access Network is noted E-UTRAN in the figure 2. Bernard, Bertrand and Alice receive the VoIP and Video flows on their smartphones. The VoIP and Video flows are transmitted in multicast over a Multimedia Broadcast Multicast Service network [10] that uses IPv4 addresses. The Multimedia Broadcast Multicast Service network is noted MBMS in the figure 2. Bénédicte, Catherine and Isabelle receive the VoIP and Video flows on their Personal Computers. The VoIP and Video flows are transmitted in multicast over a Wi-Fi network that uses IPv6 addresses. In the second stage (Figure 3), Bertrand moves. His smartphone, initially connected to the MBMS network, connects itself to the E-UTRAN network. The VoIP and Video flows are transmitted in unicast over the E-UTRAN network that uses IPv6 addresses.

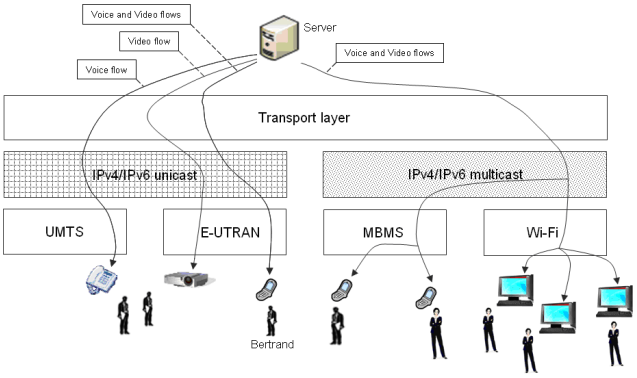


Figure 3. A scenario describing a synergy between multicast and unicast networks (second stage)

In the third stage (Figure 4), Bernard moves. His smartphone, initially connected to the MBMS network, connects itself to the Wi-Fi network. The VoIP and Video flows are transmitted in multicast over the Wi-Fi network that uses IPv6 addresses.

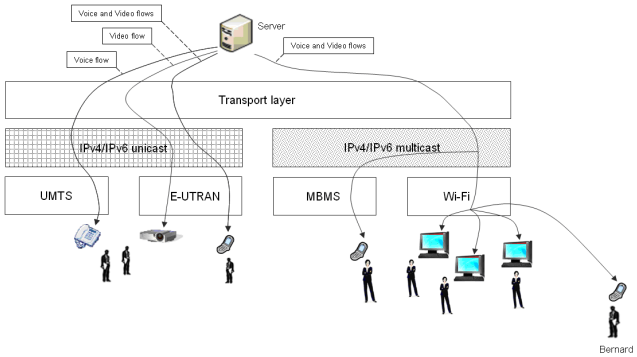


Figure 4. A scenario describing a synergy between multicast and unicast networks (third stage)

In the fourth stage (Figure 5), Bernard, who has a laptop with a Wi-Fi interface, decides to take part in the video-conference via his laptop. Having connected his laptop to the Wi-Fi network, he turns off his smartphone and takes part in the video-conference via his laptop.

 ISBN:978-1-61208-113-7

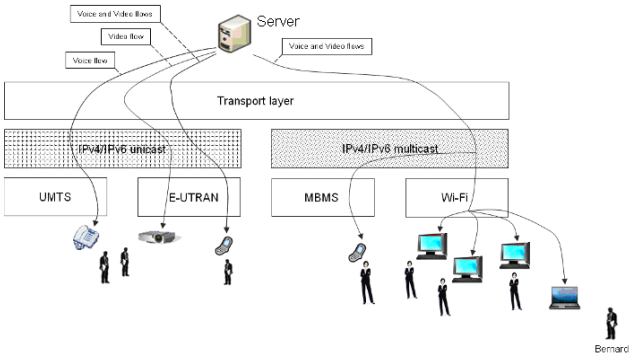


Figure 5. A scenario describing a synergy between multicast and unicast networks (fourth stage)

In the fifth stage (Figure 6), the operator who manages the MBMS network notices that the smartphone of Alice is the only device connected to this network. As the smartphone of Alice has a MBMS interface and a Wi-Fi interface, the operator decides that Alice will take part in the video-conference via the Wi-Fi interface of her smartphone without decreasing her QoS. At the end of this operation the resources of the MBMS network are no longer used.

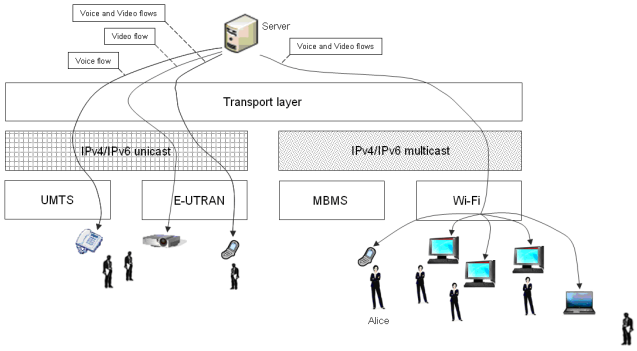


Figure 6. A scenario describing a synergy between multicast and unicast networks (fifth stage)

In the last stage of our scenario (Figure 7), Antoine, who has a laptop with a Wi-Fi interface and an E-UTRAN interface, moves. After having turned on his laptop, he receives the VoIP and Video flows via two different interfaces: the VoIP flow is received in unicast via the E-UTRAN interface, while the Video flow is received in multicast via the Wi-Fi interface.

The six stages of our scenario show that a synergy between IPv4/IPv6 unicast radio networks and IPv4/IPv6 multicast radio networks is possible and could interest a telecoms operator. But its implementation raises many questions. Here is a non-exhaustive list of questions. Knowing that the contexts of the members of the multicast group are different, in what format the server should send the data when the IP packets carrying the data are transmitted in multicast? Knowing that the structure of the multicast group may change over time, as we see by comparing the structure of the multicast group between the first and the sixth stage, how to adapt the format of the transmitted data to reflect these changes? Knowing that the users are mobile, what are the criteria for selecting a radio network during a handover?

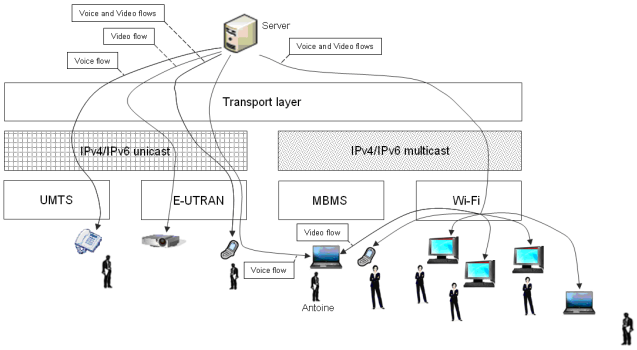


Figure 7. A scenario describing a synergy between multicast and unicast networks (sixth stage)

III. IDENTIFICATION AND ANALYSIS OF THE EXISTING WORKS

In this section we identify and analyze the works taking into account the context of the users in a multicast transmission. The first works date from the 1990s. In 1996, McCann *et al* [11] have proposed to split a multicast group according to the throughput that the terminals of the users can receive. The data flow to be transmitted is encoded in several layers, each layer being transmitted with a certain throughput. The first layer, called the base layer, is necessary to decode the flow, whereas the other layers are used to improve the QoS of the received flow. The QoS of a user is better if the number of layers decoded by his terminal is higher. Each layer is associated with a subgroup of the partition of the multicast group. The subscription to one or several subgroups is made with the Receiver-driven Layered Multicast (RLM) protocol: the terminal of a user chooses to subscribe to one or several subgroups, i.e. to receive one or several layers, according to its decoding capabilities and the available bandwidth in the network. In 2000, Yang *et al* [12] defined a utility function [13] in the terminal of each user by taking as parameter the throughput received by the terminal. The optimum of the utility function is reached when the throughput received by the terminal is equal to the throughput that it would receive if it was alone in the group. Then, having defined a utility function for each subgroup making up a partition of the multicast group, they mathematically demonstrated that there is a partition, among all the possible partitions, whose sum of the extremums of the utility functions of the subgroups making up the partition is bigger than the sum of the extremums of the utility functions of the subgroups of the other partitions. This partition, which is the optimal partition, is obtained by a centralized process. Several works redefined the utility function implemented in the terminals by Yang *et al* [12]: Maimour *et al* [14] modified the function by taking a parameter easier to get than the throughput, namely the transmission delay between the source emitting the flow of

 ISBN:978-1-61208-113-7

IP packets and each terminal; then Yousefi' zadeh *et al* [15] chose a utility function continuously differentiable to reduce the complexity of the calculations to realize to obtain the optimal partition. In 2003, Li *et al* [16] showed that the consideration of the context of the users during a transmission of a video flow in multicast requires a video source having a method of transmission capable of transmitting the flow with various throughputs. Having established a taxonomy of the various methods of transmission of a video flow, they compared the three methods (stream replication, cumulative layering, noncumulative layering) implemented in the processes of partition of a multicast group. From the years 2005, several projects aim to take into account the context of the users in the multicast architectures: in 2008, the C-Mobile project [17] defines a MBMS architecture that takes into account the context of the users; then in 2009, the C-Cast project [18] defines methods for collecting and analyzing the contexts in a multicast architecture.

In the works that we identified, a centralized process calculates the optimal partition from a utility function implemented in each terminal. However, other entities could also implement a utility function: the operators managing networks, the suppliers of a service … Each entity can have its own criteria to define an optimal partition. For example, an operator can define a single multicast group to improve the use of his networks, whereas the users will prefer to define a number of subgroups equal to the number of users for improving the consideration of the contexts of each user. When several entities take part in the partition process, what partition to choose among those proposed by every entity? Who chooses the partition?

Furthermore, in the identified works, the mobility of the users is little studied. The analysis of the mobility requires the following definition. During a handover, when a terminal disconnects itself from a network, this network is called outgoing network; when it connects itself to a new network, this network is called incoming network. The analysis of the existing studies shows two issues. During a handover, what incoming network to choose? Who chooses the network? According to Zdarsky [19], the user chooses the incoming network, while according to Antoniou *et al* [18], it is the operator managing the network that chooses the incoming network. To take into account the diverse objectives of the entities participating in the selection process, Suciu *et al* [20] proposed a method called Hierarchical and Distributed Handover (HDHO), and analyzed a scenario composed of three entities, namely a content provider, an operator managing networks, and a user. The objective of the content provider is to choose an incoming network offering a QoS adapted to the flow to be transmitted. The objective of the operator is to choose the least loaded network for transmitting the flow of the content provider. The objective of the user is to choose a network offering the best QoS/Cost ratio for receiving the flow.

It is important to notice that the mobility of a user can cause a new partition of the multicast group due to the variation of the number of the terminals connected to the outgoing and incoming networks. Conversely, a new partition can cause the mobility of one or more users as shown in the fifth stage of the scenario described in the second section. The works that we identified do not address the interactions that can exist between the partition process of a group and the selection process of an incoming network.

IV. PRESENTATION OF OUR APPROACH

In this section we present our approach to implement our three objectives. The first two objectives seem contradictory. Indeed, the more the number of members in a multicast group is higher, the better the use of network resources. But the more the number of members in a multicast group is higher, the more it will be difficult to take into account the variety of the contexts of the users. It is thus necessary to find a compromise between, on one hand, the multicast transmission of the IP packets and, on the other hand, the consideration of the contexts of the users. Our approach aims at defining several entities that will divide the multicast group into several subgroups according to the context of the users. Three entities, namely the content provider, the operator managing the networks, and the users, are involved in the partition process. The objectives of the three entities are the following ones. The objective of the content provider is to encode his content into one or more different formats according to the encoding processes available on his servers (objective 1); the objective of the operator is to transmit the data in multicast when the multicast transmission is possible and when the consumption of the resources of networks during a multicast transmission of data towards N users is lower than the consumption of the resources of networks during a unicast transmission of data towards N users (objective 2); the objective of the users is to receive the data in a format adapted to their context (objective 3).

The implementation of the third objective is made with the HDHO method. The three entities participating in the selection process of an incoming network are the content provider, the operator managing the networks, and the users. The objective of the content provider is to choose an incoming network offering a QoS adapted to the flow to be transmitted (objective 4). The objective of the operator is to choose the least loaded network for transmitting the flow of the content provider (objective 5). The objective of the user is to choose a network offering the best QoS/Cost ratio for receiving the flow (objective 6). Since our approach is based on the definition of three entities, each with their goals, we must specify how these three entities interact among them to split the multicast group and to select an incoming network during a handover. By referring to the works of Suciu *et al* [20], we chose the Free Conflict method, the Compromise and Negotiation method, and the Team Enforced method for implementing these interactions.

Since the mobility of a user can cause a new partition of the group and as a new partition can cause the mobility of one or more users, the partition process and the selection process interact among them. The interaction between the two processes is described by the heuristic presented in Figure 8. During a handover, the end of the selection process triggers the beginning of the partition process. At each iteration of the partition process, the operator managing the

 ISBN:978-1-61208-113-7

network can initiate a handover of one or more users to meet its objective. The selected networks must not degrade the QoS of the users.

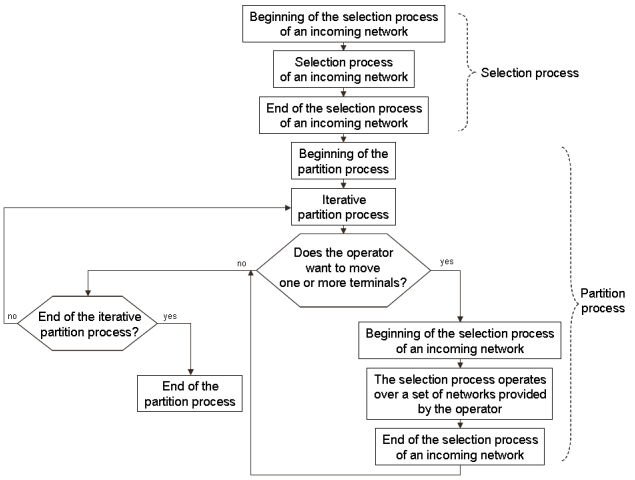


Figure 8. Heuristic describing the interaction between the selection process and the partition process

Currently we are working on the modules composing the heuristic and we reflect on the way to integrate our approach in a MBMS architecture. Figure 9 shows a possible integration. The objectives of the content provider, namely the objectives 1 and 4, can be implemented in the Broadcast Multicast Service Center (BM-SC) and in the Gateway GPRS Support Node (GGSN). The objectives of the operator, namely the objectives 2 and 5, can be implemented in the GGSN and in the Radio Network Controller (RNC). The objectives of the user, namely the objectives 3 and 6, can be implemented in the user's terminal.

V. CONCLUSION

As the optimization of the use of the resources of the networks is a major issue for telecoms operators, we initiated works that aim to, firstly, improve the use of the resources of the networks by transmitting the IP packets in multicast when it is possible, secondly, adapt the format of the data transmitted in multicast to take into account the context of the users, and thirdly preserve the Quality of Service when a user moves from a radio network towards another radio network. After having shown, through a scenario, how our work would allow to establish a synergy between multicast and unicast networks, we analyzed the works taking into account the context of the users during a multicast transmission. The analysis revealed three loopholes: the terminals of the users are the only entities that participate in the partition process; the mobility of the users is little studied; the listed works do not tackle the interactions that can exist between the partition process and the selection process. Our work, that takes into account these shortcomings, aims to, firstly, define the entities involved in the partition process and the selection process, secondly, define the objectives of these entities, thirdly, define the partition algorithm according to the objectives of each entity, fourthly, define the interaction between the partition process and the selection process. When these four steps will be made, we will model the partition process and the selection process with OPNET®.

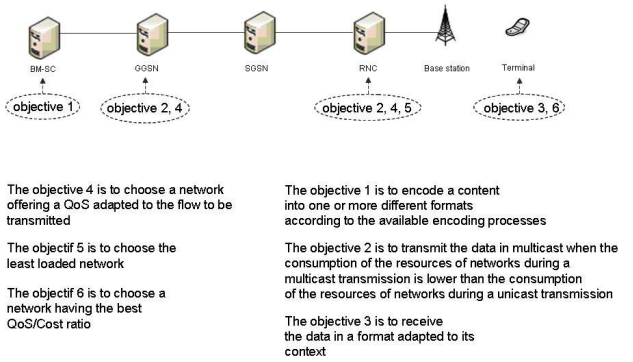


Figure 9. Integration of our approach in a MBMS architecture

REFERENCES

[1] Federal Communications Commission Spectrum Policy Task Force. Report of the Spectrum Efficiency Working Group, 2002.

[2] B.A. Fette, Cognitive Radio Technology (2nd ed). Burlington, MA: Academic Press, 2009.

[3] Federal Communications Commission. Notice of Proposed Rule Making. ET Docket no 04-113, 2004.

[4] IEEE. IEEE Draft Standard for Information technology Telecommunications and information exchange between systems Local and metropolitan area networks Specific requirements Part 22.1: Standard to Enhance Harmful Interference Protection for Low Power Licensed Devices Operating in TV Broadcast Bands, 2009.

[5] S. Deering, IETF Network Working Group: RFC 1112. Host Extensions for IP Multicasting, 1989.

[6] S. Mascolo, A. Grieco, G. Pau, M. Gerla and C. Casetti, End-to-End Bandwidth Estimation in TCP to Improve Wireless Link Utilization. European Wireless Conference, 2002.

[7] G.D. Abowd and A.K. Dey, Towards a Better Understanding of Context and Context-Awareness, LNCS 1707, pp. 304-307, 1999.

[8] H. Schwarz, G. Sullivan, T. Wiegand and M. Wien, Text of ISO/IEC 14496-10:200X/FDIS Advanced Video Coding (4th Edition). International Organisation For Standardisation ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio, 2007.

[9] P. Lescuyer and T. Lucidarme, Evolved Packet System (EPS) The LTE and SAE Evolution of 3G UMTS. Chichester, England: John Wiley & Sons, Ltd, 2008.

[10] R. Huzenlaub, MBMS from A-Z. Karlsruhe, Deutschland: INACOM GmbH, 2006.

[11] S. McCanne, V. Jacobson and M. Vetterli, Receiver-driven Layered Multicast. ACM SIGCOMM, pp. 117-130, 1996.

[12] Y.R. Yang, M.S. Kim and S.S. Lam, Optimal Partitioning of Multicast Receivers. International Conference on Network Protocol, pp. 129-140, 2000.

[13] P. Anand, Foundations of Rational Choice Under Risk (3rd ed). Oxford: Oxford University Press, 2002.

[14] M. Maimour and C. Pham, A RTT-based Partitioning Algorithm for a Multi-rate Reliable Multicast Protocol. Proceedings of the IEEE High-Speed Network and Multimedia Communications Conference, 2003.

 ISBN:978-1-61208-113-7

[15] H. Yousefi' zadeh, H. Jafarkhani and A. Habibi, Layered Media Multicast Control (LMMC): Rate Allocation and Partitioning. IEEE/ACM Transactions on Networking, 13(3), pp. 540-553, 2005.

[16] B. Li and J. Liu, Multirate Video Multicast over the Internet: An Overview. IEEE Network, pp. 24-29, 2003.

[17] J. Santos, D. Gomes, S. Sargento, R.L. Aguiar, N. Baker, M. Zafar and A. Ikram, Multicast/broadcast network convergence in next generation mobile networks. Computer Networks: The International Journal of Computer and Telecommunications Networking, 52(1), pp. 228-247, 2007.

[18] J. Antoniou, C. Christophorou, C. Janneteau, M. Kellil, S. Sargento, A. Neto, F.C. Pinto, N.F. Carapeto and J. Simoes, Architecture for Context-Aware Multiparty Delivery in Mobile Heterogeneous Networks. International Conference on Ultra Modern Telecommunications & Workshops, 2009.

[19] F.A. Zdarsky and J.B. Schmitt, Handover In Mobile Communication Networks: Who is In Control Anyway? Proceedings of the 30th EUROMICRO Conference, pp. 205-212, 2004.

[20] L. Suciu, M. Benzaid, S. Bonjour and P. Louin, Assessing the Handover Approaches for Heterogeneous Wireless Networks. 18th International Conference on Computer Communications and Networks, 2009.

 ISBN:978-1-61208-113-7

A Survey on Robust Wireless JPEG 2000 Images and Video Transmission Systems

Max Agueh
LACSC – ECE
37, Quai de Grenelle
75015 Paris, France
agueh@ece.fr

Henoc Soude
LIASD – Université Paris 8
Paris, France
soude@ai.univ-paris8.fr

***Abstract*—This paper proposes a survey of robust wireless JPEG 2000 images and video transmission systems. The performance of the presented systems is discussed and compared both in terms of time consumption and in terms of robustness against transmission errors. Some opened tracks are then discuss with a special emphasis on efficient scalable wireless JPEG 2000 image and video transmission schemes.**

Keywords- motion JPEG 2000; wireless network; video streaming; Reed-Solomon codes, Forward Error Correction.

I. INTRODUCTION

Nowadays, more and more multimedia applications integrate wireless transmission functionalities. Wireless networks are suitable for those types of applications, due to their ease of deployment and because they yield tremendous advantages in terms of mobility of User Equipment (UE). However, wireless networks are subject to a high level of transmission errors because they rely on radio waves whose characteristics are highly dependent of the transmission environment.

In wireless video transmission applications like the one considered in this paper and presented in Figure 1, effective data protection is a crucial issue.

JPEG 2000, the newest image representation standard, addresses this issue firstly by including predefined error resilient tools in his core encoding system (part 1) and going straightforward by defining in its 11th part called wireless JPEG 2000 (JPWL) a set of error resilient techniques to improve the transmission of JPEG 2000 codestreams over error-prone wireless channel

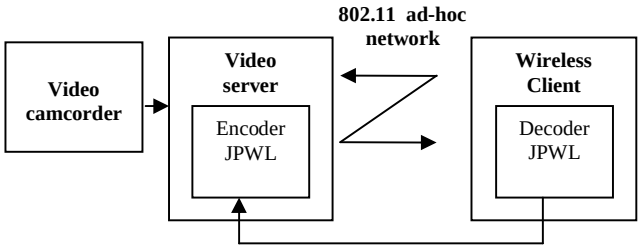


Fig. 1. Wireless video streaming system

II. OVERVIEW OF JPEG 2000 AND WIRELESS JPEG 2000

A. JPEG 2000

JPEG 2000 is the newest image compression standard completing the existing JPEG standard [1].

The interest for JPEG 2000 is growing since the Digital Cinema Initiatives (DCI) has selected JPEG 2000 for future distribution of motion pictures.

Its main characteristics are: lossy or lossless compression modes; resolution, quality and spatial scalability; transmission and progressive image reconstruction; error resilience for low bit rate mobile applications; Region Of Interest (ROI) functionality, etc.

Part 1 of the standard defines different tools allowing the decoder to detect errors in the transmitted codestream, to select the erroneous part of the code and to synchronise the decoder in order to avoid decoder crash. Even if those tools give a certain level of protection from transmission errors, they become ineffective when the transmission channel experiment high bit error rate. Wireless JPEG 2000 (JPEG 2000 11th part) addressed this issue by defining techniques to make JPEG 2000 codestream more resilient to transmissions errors in wireless systems.

B. Wireless JPEG 2000 (JPWL)

Wireless JPEG (JPWL) specifies error resilience tools such as Forward Error correction (FEC), interleaving, unequal error protection.

In this paper we present a wireless JPEG 2000 video streaming system based on the recommendations of JPWL final draft [2].

In [3], the description of the JPWL system is presented and the performance of its Error Protection Block (EPB) is evaluated. A fully JPEG 2000 Part 1 compliant backward compatible error protection scheme is proposed in [4]. A memoryless Binary Symmetric Channel (BSC) is used for simulations both in [4] and [3]. However, as packets errors mainly occur in bursts, the channel model considered in those works is not realistic. Moreover JPEG 2000 codestreams interleaving is not considered in [4].

In this paper we address the problem of robust and efficient JPEG 2000 images and video transmission over wireless networks. The paper is organized as follows: In section II, we

 ISBN:978-1-61208-113-7

present a state of art of wireless JPEG 2000 multimedia communication systems along with the challenges to overcome in terms of codestreams protection against transmission errors. In section III, we provide an overview of channel coding techniques for efficient JPEG 2000 based multimedia networking. Finally section IV, provides discussions and prospective issues for future distribution of motion JPEG 2000 images and video over wireless networks.

III. WIRELESS JPEG 2000 MULTIMEDIA COMMUNICATION SYSTEM AND ITS CHALLENGES

In high error rate environments such as wireless channels, data protection is mandatory for efficient transmission of images and video. In this context, Wireless JPEG 2000 the 11th part of JPEG 2000 [2] uses different techniques such as data interleaving, Forward Error Correction (FEC) with Reed-Solomon (RS) codes etc. in order to enhance the protection of JPEG 2000 codestreams against transmission errors.

In wireless multimedia system such as the one considered in this paper (see Figure 1), a straightforward FEC methodology is applying FEC uniformly over the entire stream (Equal Error Correction - EEP). However, for hierarchical codes such as JPEG 2000, Unequal Error

Protection (UEP) which assigns different FEC to different portion of codestream has been considered as a suitable protection scheme.

Since wireless channels' characteristics depend on the transmission environment, the packet loss rate in the system also changes dynamically. Thus a priori FEC rate allocation schemes such as the one proposed in [5] are less efficient. Two families of data protection schemes address this issue by taking the wireless channel characteristics into account in order to dynamically assign the FEC rate for JPEG 2000 based images/video. The first family is based on a dynamic layer-oriented unequal error protection methodology whereas the second relies on a dynamic packet-oriented unequal error protection methodology. Hence, in the first case, powerful RS codes are assigned to most important layers and less robust codes are used for the protection of less important layers. It is worth noting that in this case, all the JPEG 2000 packets belonging to the same layer are protected with the same selected RS code. Examples of layer-oriented FEC rate allocation schemes are available in [6] and [7]. On the other side, in packet-oriented FEC rate allocation schemes such as the one presented in [7], RS codes are assigned by decreasing order of packets importance. In [7], we demonstrate that the proposed optimal packet-oriented FEC rate allocation is more efficient than the layer-oriented FEC rate allocation scheme presented in [6] and [7]. However, layer-based FEC rate allocation schemes have low complexity while packet-oriented FEC allocation methodologies are complex especially when the number of packets in the codestream is high. In this case, packet oriented FEC schemes are unpractical for highly time-constrained images/video streaming applications. In this case switching to a layer oriented FEC rate allocation scheme is more interesting.

The smart FEC rate allocation scheme proposed in [9] address this issue by allowing switching from a packet oriented FEC scheme to a layer oriented scheme such as the ones proposed in [10].

In section III.A we present the packet oriented system proposed in [7] to address the issue of robust JPEG 2000 images and video transmission over wireless network. Then in section III.B the layer-oriented scheme proposed in [10] is described. Finally, in section III.C we present the system proposed in [9] to unify packet and layer based scheme.

A. *Optimal Packet-oriented FEC rate allocation scheme for robust Wireless JPEG 2000 based multimedia transmission*

The functionalities of the proposed JPWL packet-oriented system are presented in Figure 2 The aim of this system is to efficiently transmit a Motion JPEG 2000 (MJ2) video sequence through MANET channel traces.

The system is described as follows:

The input of the JPWL codec is a Motion JPEG 2000 (MJ2) file. The JPEG 2000 codestreams included in the MJ2 file are extracted and indexed.

These indexed codestreams are transmitted to the JPWL encoder ([2] presents a more accurate description of the used JPWL encoder) which applies FEC at the specified rate and adds the JPWL markers in order to make the codestream compliant to Wireless JPEG 2000 standard. At this stage, frames are still JPEG 2000 part 1 compliant, which means that any JPEG 2000 decoder is able to decode them.

To increase JPWL frames robustness, an interleaving mechanism is processed before each frame transmission through the error-prone channel. This is a recommended mechanism for transmission over wireless channel where errors occur in burst (contiguous long sequence of errors). Thanks to interleaving the correlation between error sequences is reduced.

The interleaving step is followed by RTP packetization. In this process, JPEG 2000 codestream data and other types of data are integrated into RTP packets as described in [11].

 ISBN:978-1-61208-113-7

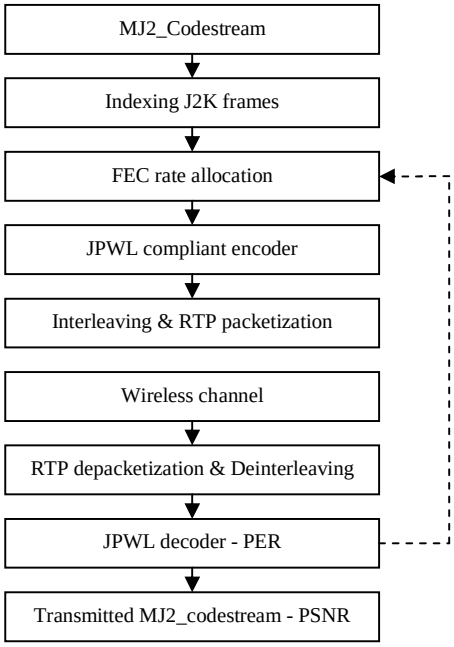


Fig. 2. JPWL based system functionalities

RTP packets are then transmitted through the wireless channel which is modelled in this work by a Gilbert channel model. At the decoder side, after depacketization, the JPWL decoder corrects and decodes the received JPWL codestreams and rebuilds the JPEG 2000 frames. At this stage, parameters such as Packet Error Rate (*PER*) are extracted, increasing the knowledge of the channel state. The decoder sends extracted parameters back to the JPWL encoder via the Up link. The last process of the transmission chain is the comparison between the transmitted and the decoded image/video. Figure 3 presents JPEG 2000 codestreams transmission through the JPWL packet-oriented FEC system

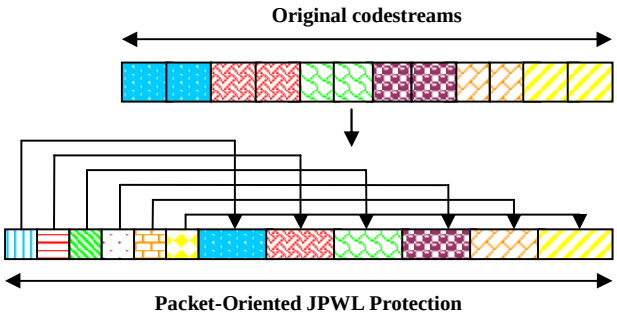


Fig. 3. JPEG 2000 codestreams transmission through the JPWL packet-oriented FEC system

B. *Optimal Layer-oriented FEC rate allocation scheme for robust Wireless JPEG 2000 based multimedia transmission*

Unlike the system described in [7], where the FEC rate allocation scheme is packet oriented, in the current system we consider a layer oriented FEC rate allocation scheme. In other words the difference between both systems is the FEC rate allocation module. Actually, in the packet oriented scheme the redundancy is added by taking the packets importance into account (see Figure 3) while in the layer oriented scheme we rely on layers importance to allocate the adequate RS codes (see Figure 4).

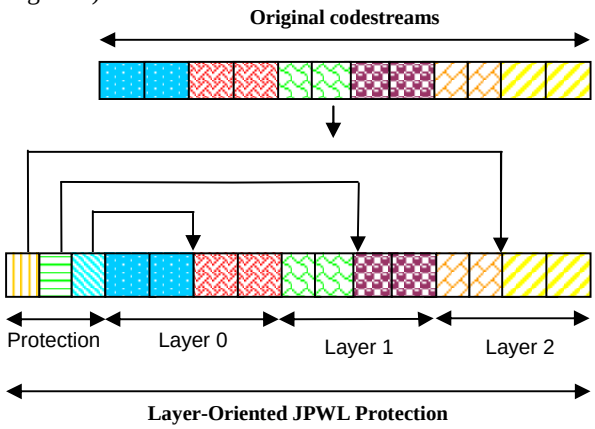


Fig. 4. A JPEG 2000 codestreams transmission through the JPWL layer-oriented FEC system

C. *Smart combined Packet/layer based FEC rate allocation scheme for robust Wireless JPEG 2000 based multimedia transmission*

The functionalities of the proposed smart JPWL based system are presented in Figure 5.

In this system, indexed JPEG 2000 codestreams are transmitted to the smart FEC rate allocation module. If the number of data packets available in the codestreams is low (typically under the defined smart threshold), the smart module uses the optimal packet-oriented FEC rate allocation methodology presented in [7] whereas it switches to the dynamic layer-oriented FEC rate allocation methodology presented in [10] when the number of data packets is high. Ones the protection rate determined, the codestreams are transmitted to the JPWL encoder which applies FEC at the specified rate and adds the JPWL markers in order to make the codestream compliant to Wireless JPEG 2000 standard. Hence, Figures 3 and 4 correspond to the JPWL protection where redundant data are added to original codestreams. If the JPEG 2000 Frame which is being processed is constituted by less than a defined threshold (*smart_thresh*) , then the smart FEC rate allocation scheme emulates a scenario similar to the one presented in Figure 3 (packet-oriented FEC rate allocation). Otherwise, it emulates the scenario of Figure 4 (dynamic layer-oriented FEC rate allocation). Protected data are then interleaved and transmitted.

 ISBN:978-1-61208-113-7

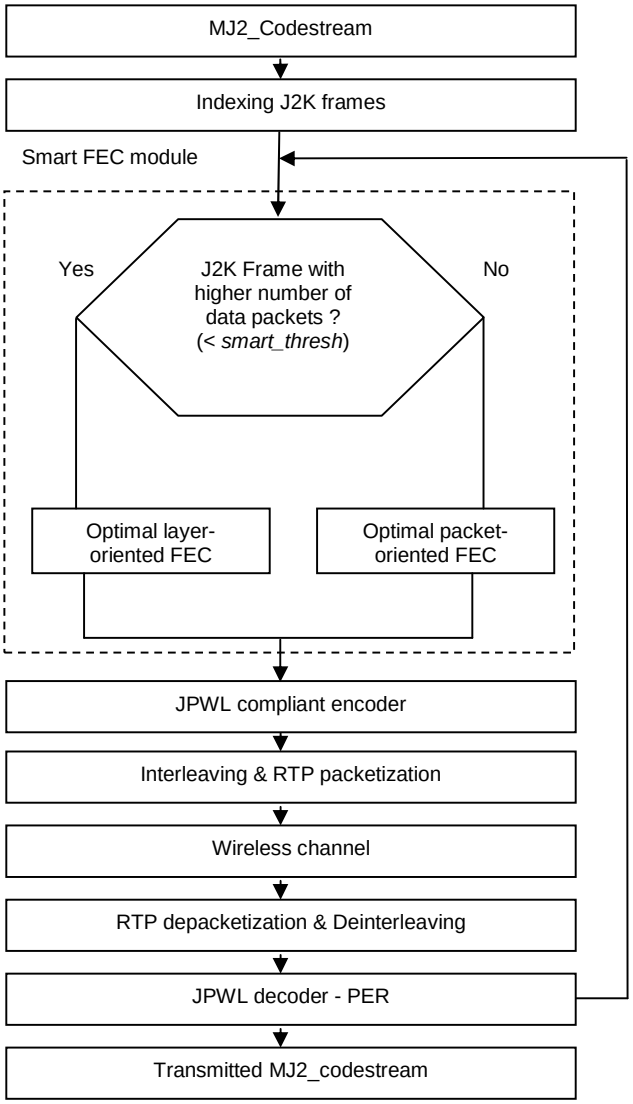


Fig. 5. JPEG 2000 transmission over the smart JPWL system

The interest of the smart FEC rate allocation scheme is to allow switching from the scenario presented in Figure 3 to the scenario described in Figure 4, reducing by this way the complexity of the FEC rate allocation process. Hence, in case of highly layered images/video streaming, the time needed to select the suited FEC rate is significantly reduced. In the following section we present the packet-oriented and layer-oriented algorithm considered in this paper.

IV. RESULTS

A. Performance of layer based FEC scheme in terms of time consumption

In Figure 6 the run time of the proposed layer based FEC rate allocation scheme is plotted versus the number of data packets available in the JPEG 2000 codestreams. This curve is compared to the one achieved using the optimal packet oriented FEC rate allocation scheme [7]. These results are achieved using an Intel core Duo CPU 2.9 Ghz Workstation.

As packet-oriented and layer oriented schemes are linked by the number of layers available in each image, we vary this parameter in order to derive some comparable results. In the considered scenario, the number of available resolution and component of JPEG 2000 frames are fixed (resolution = 10 and component = 1) because these parameters do not impact the time-performance of layer oriented FEC rate allocation schemes. In Figure 6 each packet (i) corresponds to a specific JPEG 2000 frame (with a specific quality layer).

In this scenario, the available bandwidth in the system is set to 18 Mbits/s ($B_{av} = 18\ Mbits/s$). It is worth noting that in practice few existing JPEG 2000 codecs allow high quality scalability and to our knowledge, none of them can handle more than 50 quality layers. Hence, the considered scenario allows generalization to future high quality layer scalable FEC rate allocation systems.

In Figure 6 we notice that both layer and packet oriented scheme have a run time linearly increasing with the number of packets available in the codestreams. However, the optimal layer based FEC scheme is significantly less time consuming than the packet based FEC scheme. For codestreams containing less than 1000 packets (quality layers $\leq$ 10) the packet oriented FEC scheme is 3 times more time consuming than our optimal layer based FEC scheme. For JPEG 2000 codestreams, whose number of packets is between 1000 and 5000 (quality layers between 10 and 50) the packet oriented scheme is up to 5 times the run time of the layer based FEC scheme. Since existing JPEG 2000 codecs handle less than 50 quality layers, our proposed optimal layer based scheme is a good candidate for real-time JPEG 2000 codestreams over wireless channel as its yields low time consumption.

 ISBN:978-1-61208-113-7

The proposed optimal layer based scheme, due to its low time consumption, could be viewed as a good candidate for future high quality layer scalable wireless JPEG 2000 based images and video streaming applications.

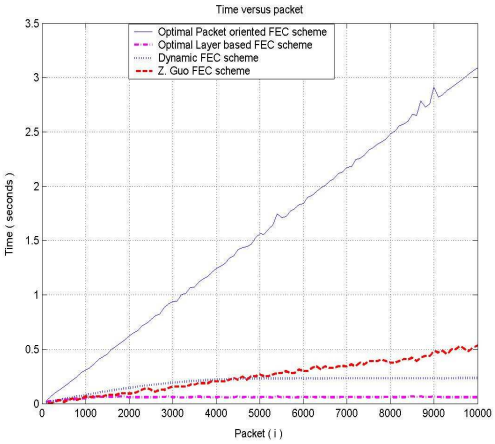


Fig.6 Time versus packets: Fixed image resolution (R=10) –Varying quality layers (0 to 100) – One component (C =1)

Although the layer based scheme achieves good performances in terms of time consumption in comparison to packet oriented FEC rate allocation schemes, the last ones present better performance in terms of visualization quality especially for highly noisy channels. In the following section we demonstrate the effectiveness of the optimal layer based FEC scheme thanks to a client/server application of Motion JPEG 2000 video streaming over real ad-hoc network traces.

B. Packet-oriented and Layer-oriented FEC rate allocation for Motion JPEG 2000 video streaming over real ad-hoc network traces

In this section we present the results achieved while streaming Motion JPEG 2000 based video over real ad-hoc network channel traces [13][13] and we demonstrate that the proposed optimal layer based scheme outperforms existing layer oriented FEC schemes even if for highly noisy channel it is less efficient than packet oriented FEC scheme. The comparison is handled both in terms of Structural Similarity (SSIM) [14] and in terms of successful decoding rate. We derive the Mean SSIM metric of the Motion JPEG 2000 video sequence by averaging the SSIM metrics of the JPEG 2000 images contained on the considered video sequence. It is worth noting that each SSIM measure derived is associated to a successful decoding rate metric which corresponds to decoder crash avoidance on the basis of 1000 transmission trials.

The considered wireless channel traces are available in [13] and the video sequence used is *speedway.mj2* [12] containing 200 JPEG 2000 frames generated with an overall compression ratio of 20 for the base layer, 10 for the second layer and 5 for the third layer. Figure 7 presents the successful decoding rate of the motion JPEG 2000 video sequence *speedway.mj2* [12] transmission over real ad-hoc network channel traces [13]. We observe that for highly noisy channels ($C/N \le 15\,dB$), the proposed optimal layer outperforms other layer based FEC schemes but is less efficient than the packet oriented scheme. For noisy channel ($15\,dB \le C/N \le 18\,dB$), we notice that all layer based UEP schemes exhibit similar performances in terms of successful decoding rate.

For low noisy channel ($C/N \ge 18\,dB$) all the FEC schemes yield the same improvement in terms of successful decoding rate.

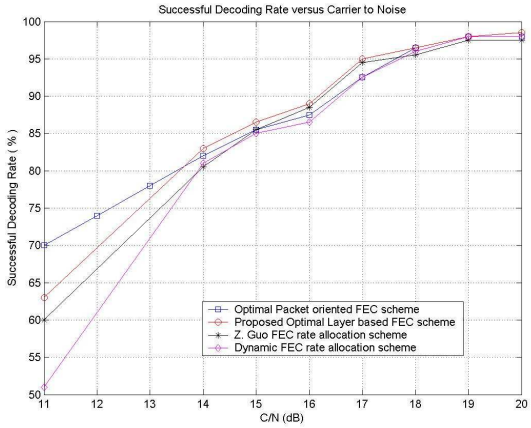


Fig.7 Successful decoding rate versus Carrier to Noise Ratio

In Figure 8 we show that our proposed optimal layer based FEC rate allocation scheme still outperforms other layer based schemes in terms of Mean SSIM. This is due to the fact that the base layer which is the most important part of the codestreams is highly protected in our proposed scheme, in comparison to other layer based schemes, guaranteeing this way a good quality for the visualization.

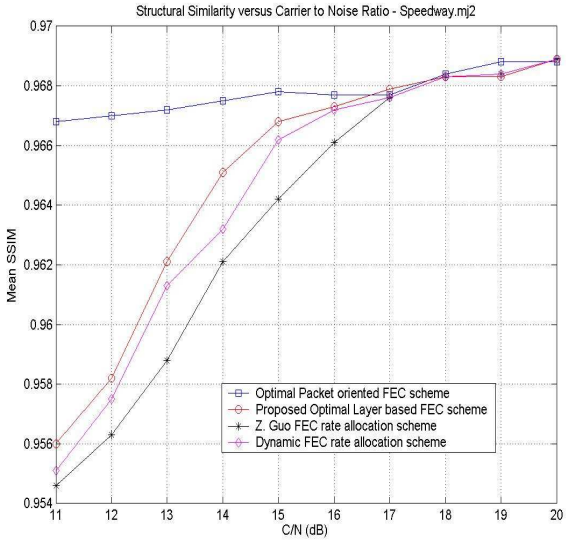


Fig.8 Mean Structural Similarity versus Carrier to Noise Ratio

 ISBN:978-1-61208-113-7

V. CONCLUSION AND OPEN ISSUE : SCALABLE JPEG 2000 TRANSMISSION

Many problems are still to be addressed in the framework of JPEG 2000 codestreams transmission over wireless networks. Image scalability based on dynamic available bandwidth estimation is one of those problems. In the literature, proposed image scalable systems have been implemented using a fixed available bandwidth in their considered scenarios [15], [16]. This assumption is no longer true in wireless systems because they rely on radio waves whose characteristics depend on the transmission environment. Moreover, few of the proposed systems addressed simultaneously the bandwidth estimation problem and the issue of smoothness for JPEG 2000 codestreams scalability. In [17], we addresses both issues by proposing a scalable and non aggressive wireless JPEG 2000 image and video transmission algorithm based on a dynamic bandwidth estimation tool.

The main limitation of the scalable system proposed in [17] is that it handles only one wireless client. However, this limitation could be overcome by generalizing the proposed algorithm to multiple wireless clients' scenario. We propose a framework for this generalization which opens the path for efficient wireless JPEG 2000 codestreams transmission in Next Generation Networks which are characterize by the cohabitation of multiples wireless devices having different standards requirements and different capacities.

REFERENCES

[1] D.S. Taubman and M.W. Marcellin, " JPEG 2000 Image Compression Fundamentals, Standards and Practice," In: Kluwer Academic Publishers, The Netherlands 2001, 2001

[2] JPWL, (2005) JPEG 2000 part 11 Final Draft International Standard, ISO/IEC JTC 1/SC 29/WG 1 N3797

[3] F. Dufaux and D. Nicholson, "JPWL: JPEG 2000 for Wireless Applications," *Proceeding of SPIE* -- Volume 5558 - Applications of Digital Image Processing XXVII, Andrew G. Tescher, Editor, pp. 309-318, November 2004

[4] D. Nicholson, C. Lamy-Bergot, , Naturel, and C. Poulliat, " JPEG 2000 backward compatible error protection with Reed-Solomon codes," *IEEE Transactions on Consumer Electronics*, vol. 49, n. 4, pp.855-860, Nov. 2003

[5] M. Agueh, F.O. Devaux and J.F. Diouris, "A Wireless Motion JPEG 2000 video streaming scheme with a priori channel coding," *Proceeding of 13th European Wireless 2007* (EW-2007), April 2007, Paris France

[6] Z. Guo, Y. Nishikawa, R. Y. Omaki, T. Onoye and I. Shirakawa, "A Low-Complexity FEC Assignment Scheme for Motion JPEG 2000 over Wireless Network". *IEEE Transactions on Consumer Electronics*, Vol. 52, Issue 1, Feb. 2006 Page(s): 81 – 86

[7] M. Agueh, J.F. Diouris, M. Diop, F.O. Devaux, "Dynamic channel coding for efficient Motion JPEG 2000 streaming over MANET," *Proceeding of Mobimedia2007*, August 2007, Nafpaktos, Greece

[8] M. Agueh, J.F. Diouris, M. Diop, F.O. Devaux, C. De Vleeschouwer and B. Macq, "Optimal JPWL Forward Error Correction rate allocation for robust JPEG 2000 images and video streaming over Mobile Ad-hoc Networks," *EURASIP Journal on Advances in Signal* Proc., Spec. Issue wireless video, Vol. 2008, Article ID 192984, doi: 10.1155/2008/192984

[9] Agueh, M., Ataman, S. & Henoc, S. (2009, a). A low time-consuming smart FEC rate allocation scheme for robust wireless JPEG 2000 images and video transmission. *Proceeding of Chinacom2009*, 2009, Xi'an, China

[10] M. Agueh and S. Henoc, "Optimal Layer-Based Unequal Error Protection for Robust JPEG 2000 Images and Video Transmission over Wireless Channels," *Proceeding of MMEDIA2009*, pp.104- 109, 2009 First International Conference on Advances in Multimedia, 2009

[11] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, " RTP: A Transport Protocol for Real-Time Applications," STD 64, RFC 3550, July 2003.

[12] Speedway video sequences have been generated by UCL.
Available: http://euterpe.tele.ucl.ac.be/WCAM/public/Speedway%20Se quence/

[13] Loss patterns acquired during the WCAM Annecy 2004 measurement campaigns IST-2003-507204 WCAM, Wireless Cameras and Audio-Visual Seamless Networking.
project website: http://www.ist-wcam.org

[14] Z. Wang, A.C. Bovik, H. R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions. on Image Processing*., vol. 13,no. 4, pp 600-612, Apr. 2004

[15] M. Li, and C. Chang, "A two-way available bandwidth estimation scheme for multimedia streaming networks adopting scalable video coding," *Proceeding of IEEE Sarnoff Symposium*, p1-p11, Princeton, USA, 2009.

[16] F. Devaux, J. Meessen, C. Parisot, J. Delaigle, B. Macq, and C. De Vleeschouwer, "A flexible video transmission system based on JPEG 2000 conditional replenishment with multiple reference," Proceeding of IEEE ICASSP 2007, Honolulu 2007, USA

[17] C. Mairal and M. Agueh**,** " Smooth and Scalable Wireless JPEG 2000 images and video streaming with dynamic Bandwidth Estimation," *Proceeding Of the Second International conference on advances in Multimedia*, June 2010, Athens, Greece

 ISBN:978-1-61208-113-7

Analysis of Reliable and Scalable Video-On-Demand Networks

Nader F. Mir
Dept. of Electrical Engineering, San Jose State University
One Washington Square
San Jose, California 95192, U.S.A
Office Phone: (408) 924-3986
Email: nader.mir@sjsu.edu

Savitha Ravikrishnan and
Meera M. Nataraja,
Dept. of Electrical Engineering, San Jose State University
One Washington Square
San Jose, California 95192, U.S.A
Email: savy8429@gmail.com
Email: mechins@gmail.com

***Abstract*-Various architectures have been proposed and implemented to handle the rapid growth in demand for video delivering technologies. This paper implements and thoroughly examines various Video-on-Demand (VoD) architectures using the NS2 simulation tool. The simulation tool used for the project provides an efficient platform to analyze different architectures and obtain performance metrics and plots. This article analyzes the VoD traffic, compares the performance of VoD architectures under different network conditions and suggests efficient network parameters required for VoD architectures to become more reliable and scalable.**

Keywords-Video On Demand; Interactive TV; Tele-conferencing; VoD Architecture.

I. INTRODUCTION

Video-on-demand (VoD) lso provides interactivity, large catalogue of content and flexibility to watch content at the user's leisure rather than being bound by time limits. Higher network bandwidth speeds, faster CPU's and mobile internet such as Wifi and 3G have fueled the advancement of video-on-demand technologies. Video-on-Demand has wide range of applications in the field of entertainment, education, and business. Some of the examples of the applications are Movies on demand, Interactive video games, Interactive news television, Distance learning, Catalogue browsing, Interactive advertising, Video-conferencing, etc [1]. Video-on-Demand systems are continuously evolving and there have been contributions from researchers and the industry to provide varied capabilities and improved architectures. We analyze some of the architectures behind the Video-on-Demand systems. We use multiple approaches to provide this analysis, measure various parameters and give graphical representation of the results.

Performance of the Video-on-Demand system can be measured by evaluating various metrics. The transport metrics [2] measured in this paper are *packet loss*, *packet delay*, and *jitter*.

Packet loss in VoD system can be caused due to various reasons like bandwidth limitations, network congestion, link failures, transmission errors, signal degradation over the network medium, corrupted packets and faulty hardware. With UDP based video streaming protocols, a loss of packets will affect the video streams as information cannot be recovered as no retransmissions occur unless the upper layer protocol has support for it. In case of TCP based protocols, retransmissions make sure that data is somehow sent to the client, but retransmissions can induce delays thereby causing frozen images.

Packet delays are very common in packet-based networks. The various possible routes the packet may have to travel and various factors like hardware, bandwidth speed and congestion in the different routes can cause a delay in the packet arrival. Usually video transmission protocols handle the arrival of delayed packets through buffering. When the delay of arriving packets exceeds the buffer size the packet is dropped. This drop can affect video quality.

Jitter is defined as a measure of the variability over time of the packet latency of a network. A short-term variation in the packet arrival time can be caused due to network congestion, difference in routes and hardware errors. Usually a small jitter buffer is present in the client side to smooth out the variations by collecting out of order frames and sequencing it in the correct order. With severe jitter, the buffer may overflow causing distorted video.

The remaining of this article focuses on the performance evaluation of the video-on-demand systems with centralized architectures examining various clients attached to the system.

II. PERFORMANCE EVALUATION

This paper implements VoD architectures in Network Simulator 2 [3]. Simulation environment used in the project is MyEvalvid_RTP Framework supported on Network Simulator 2 [4]. Data analysis implementation is done to calculate the delay, jitter, inter-packet delay. The implementation is done

 ISBN:978-1-61208-113-7

according to RFC3550 [5]. The delay of particular packet 'i' has been calculated using the formula provided below.

Delay (i) = Receive_time (i) – Send_time (i) (1)

Average delay has been calculated using:

Average delay per client = $[\sum_{k=1}^{N} (delay)_k] / N$ (2)

where N is the number of clients in the system. The interpacket delay between two successive packets 'i' and 'j' has been calculated using the formula provided below.

Interpacket_delay(i,j) = [Receive_time(j)-Receive_time(i)]-[Send_time(j)-Send_time(i)] (3)

The average interpacket delay of VoD network has been calculated using the equation provided below.

Average interpacket delay per client = $[\sum_{k=1}^{N}$ (interpacket delay)$_k$] / N (4)

Jitter is defined as packet delay variation. In particular, inter-arrival jitter is defined as mean deviation of interpacket-delay and has been implemented according to RFC3550 [10].

Jitter(i) = Jitter(i-1) + [|Interpacket_delay(i-1,i)| - Jitter(i-1)] / 16 (5)

Average jitter has been calculated using the equation provided below.

Average jitter per client = $[\sum_{k=1}^{N} (jitter)_k] / N$ (6)

Now, we compare two architectures: Centralized and Content-based networks.

III. CENTRALIZED ARCHITECTURES

Centralized architecture is one in which the central video-on-demand server's are placed in the core of the network. The various Internet service providers (ISPs) are connected to the core network through access networks which may take multiple router hops to reach the VOD server. Figure 1 shows the simulated centralized architecture with 5 clients. Figure 2 shows the flow diagram for the centralized architecture. The clients in turn are connected to the ISPs. To better understand the various factors affecting video quality at client side, a network as shown below was chosen. The VOD server is connected to the router with a 30Mbps link and a delay of 10ms. The rest of the routers from 1 to 16 are connected with a constant bandwidth of 10Mbps and varying delays.

The clients were in-turn connected to their respective ISP's with different link speeds to simulate the currently popularly available client bandwidth rates. Clients 1-5 were connected at the rates of 768 Kbps, 1024 Kbps, 2 Mbps, 4 Mbps, and 6 Mbps respectively. All clients were simulated to request a same video of length 750 frames. Various factors such as delay, inter-packet delay and jitter have been studied

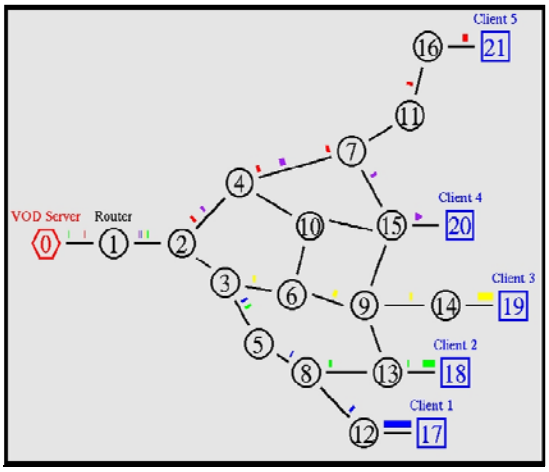


Figure 1. Simulated Centralized Architecture with 5 clients

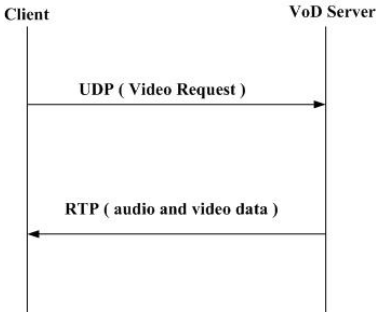


Figure 2. Flow diagram for Centralized Architecture

Client 1, which was connected at 768 Kbps had the most varying data for delay. Large variance in delay can affect the video quality badly. In case of client 2 connected at 1024 Kbps the delay values are significantly stable. The reason for such a variance between client-1 and client-2 is the bit rate at which the video was encoded, which is 1024 Kbps. Clients 3, 4 and 5 show a constant delay data.

An interesting thing to note about clients 3, 4 and 5 is that even with better bandwidth client 5 has higher delay than client 4 and client 3. This is due to the delay between the different routers and the higher number of hops that a frame takes to reach client 5, when compared to client 4. Client 3 has same number of hops as client 5, but the delay on a per hop basis has been configured to be less for client 3 which has actually boosted its performance.

The smoothened absolute value of inter frame delay gives the final jitter. Large variance in inter frame delay depicts out

 ISBN:978-1-61208-113-7

of order frames, which can lead to very bad video quality. The client bandwidth plays a major role in the inter frame delay. From the above simulation, client 1 and client 2 which have the least bandwidth connectivity's to their respective ISPs have the most variance in inter frame delay. With higher bandwidth connectivity clients 3, 4 and 5 have very less variance. The inter frame delay plays a major role in the final inter frame jitter values. With a large variance in delay, the jitter effect is more pronounced in clients 1 and 2. Jitter corresponds to choppiness in the video and with high jitter values, video quality is drastically affected. Figures 3, shows the delay for video transmission from VoD Server to Client.

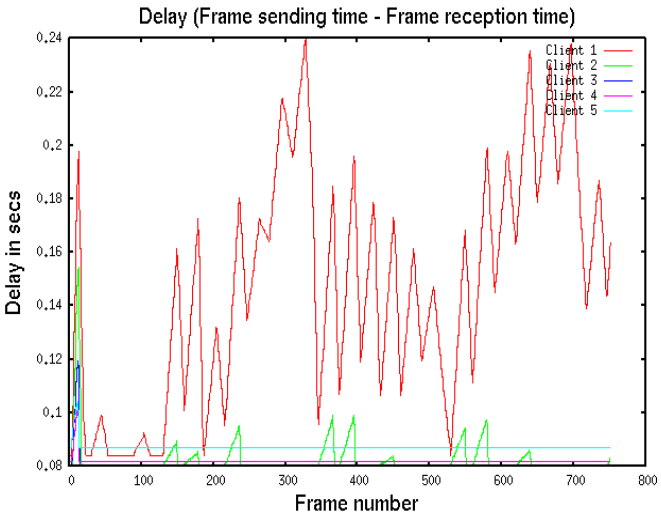


Figure 3. Delay for video transmission from VoD Server to Client

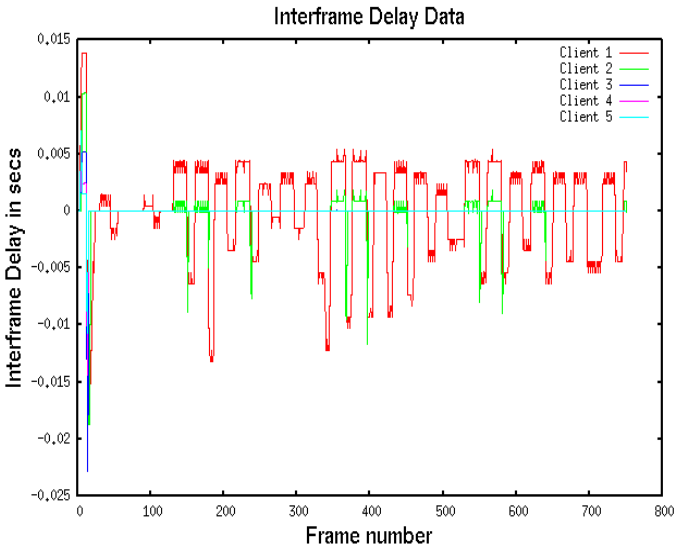


Figure 4. Centralized Architecture Interpacket delay plot

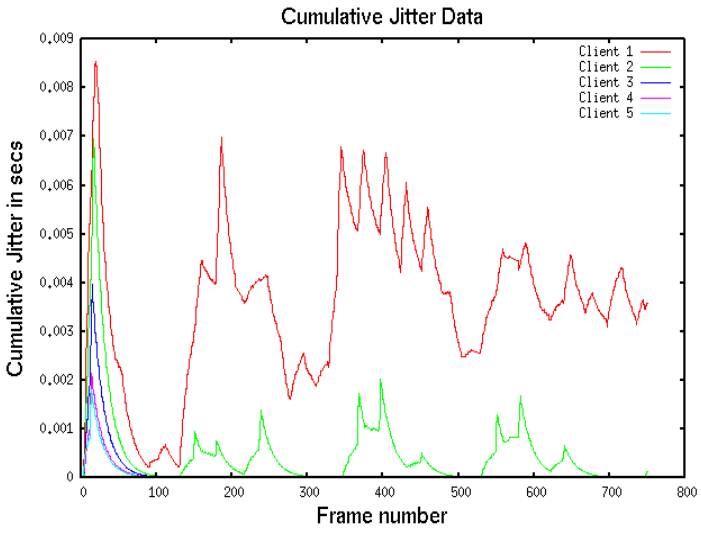


Figure 5. Centralized Architecture Jitter plot

A. Case 1

The effect of increasing number of clients in centralized architecture is provided below. Video on demand is currently in a high growth spurt. Scalability of architecture plays a vital role in its selection for widespread use. To better understand the effect of increasing number of clients on centralized architecture, all clients are connected with a constant bandwidth link of 2 Mbps. All router nodes are interconnected with a 10 Mbps link. The number of clients is continuously varied and various factors like average delay, jitter and packet loss are studied.

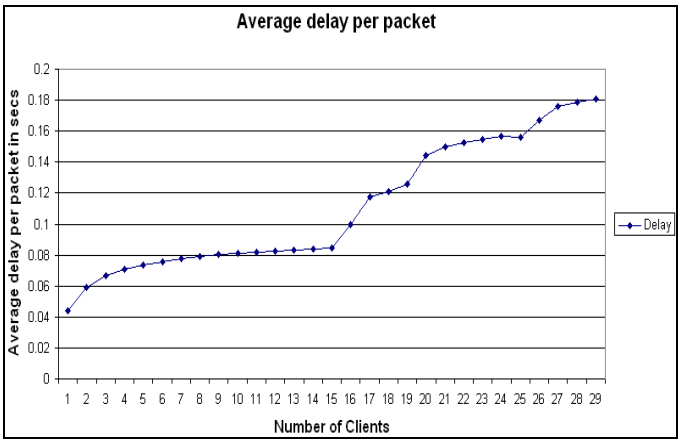


Figure 6. Average delay with the increase in the number of clients

All clients are configured to request the same video in our analysis, which gives consistent values for the graphs. As we can see in Figure 6, with increasing number of clients choking

the bandwidth, the average delay per packet increases. In the scaled down network we can clearly see that till 16 simultaneous clients, the variance in delay is very smooth and levels down. When the 16 client threshold is reached a drastic increase in delay is seen.

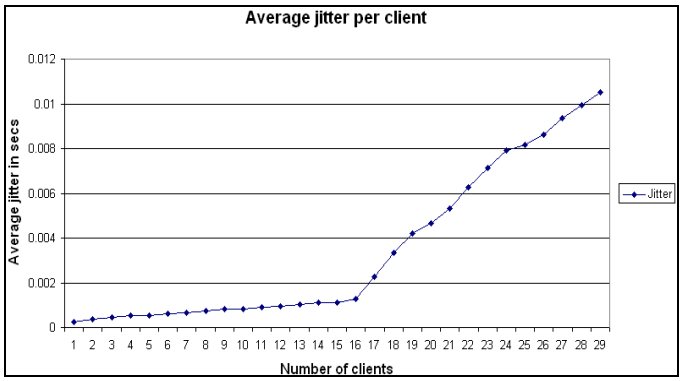


Figure 7. Average jitter with increase in the number of clients

The average jitter per client shown in Figure 7 is calculated by averaging the inter frame jitter on a per client basis. The effect of delay on Jitter is clearly seen in the above graph. When the delay was constant and under 0.1 seconds till 16 clients, the jitter was varying in a constant manner. Once the sweet spot of 16 clients is exceeded a drastic change in jitter is seen.

Packet loss percentage is calculated as

100 x (1 – Packets seen / Packets expected)

Packet loss shown in Figure 8 shows the most drastic effect on video quality. When frames of video data are lost, video starts stuttering and become distorted. The drastic effect of packet loss with increasing number of clients once it reaches a particular count shows one of the major disadvantages of the centralized architecture. The video quality has a very high dependency on the access network links.

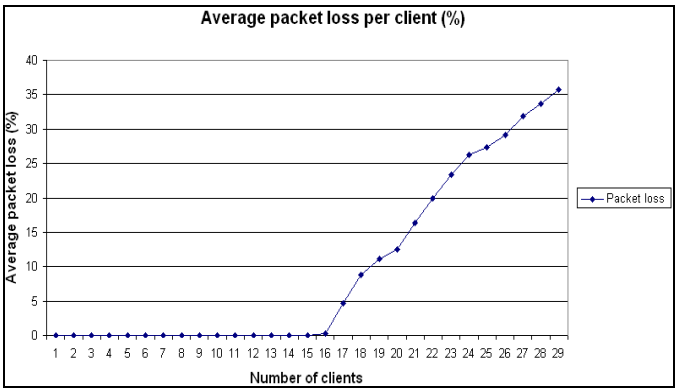


Figure 8. Average packet loss with increase in the number of clients

B. Case 2

The effect of increasing access network bandwidth is provided below. From the previous analysis with 5 clients, factors such as client bandwidth, router bandwidth and delay were shown as major factors affecting video transmission. In this current experiment the client bandwidth is taken to be constant value of 2 Mbps, which is sufficient to transmit a 1024 Kbps encoded video. The major effect on quality of video transmission in case of centralized architecture is due to the bandwidth between the access links connecting the core of the network to the edge of the network.

As access network bandwidth plays a major role in the video quality transmitted to the client, the same test as above has been conducted with a constant number of clients but varying bandwidth. The number of clients connected to the centralized video on demand server is kept at a constant value of 20. The bandwidth across the access network between each node is constantly varied from 5 Mbps to 20 Mbps and the effect of varying bandwidth is studied.

The average delay per packet has only been calculated for the packets that have been successfully received on the client end. With this data it is clear that increasing bandwidth reduces delay considerably. Once we reach 16 Mbps, the delay becomes constant and is now dependent on the delay between links of the access network. A reduction in average delay corresponds to reduction in jitter. Similar to delay, when sufficient bandwidth is present, the variation in delay becomes minimal. Figures 9, 10 and 11 are the results for this case.

 ISBN:978-1-61208-113-7

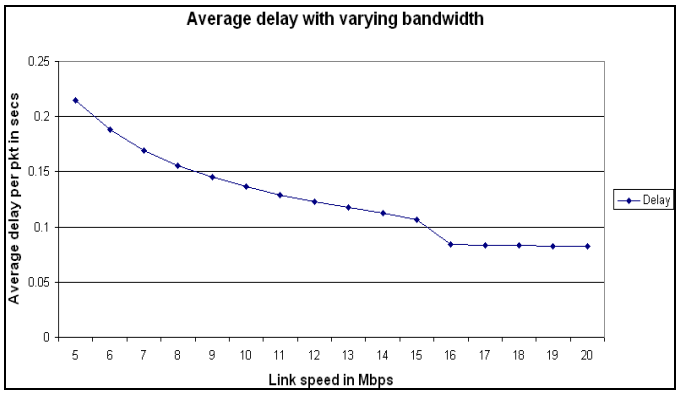


Figure 9. Average delay with increase in bandwidth

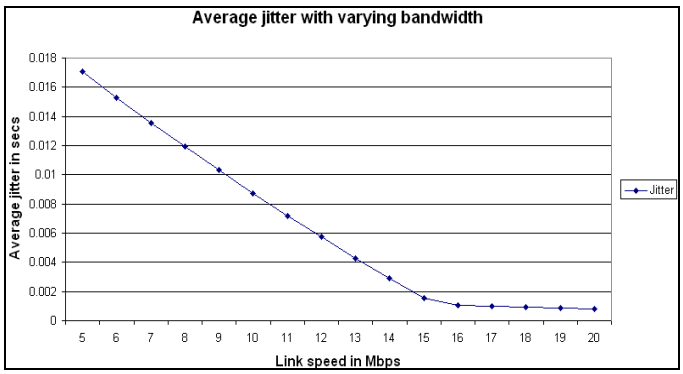


Figure 10. Average jitter with increase in bandwidth

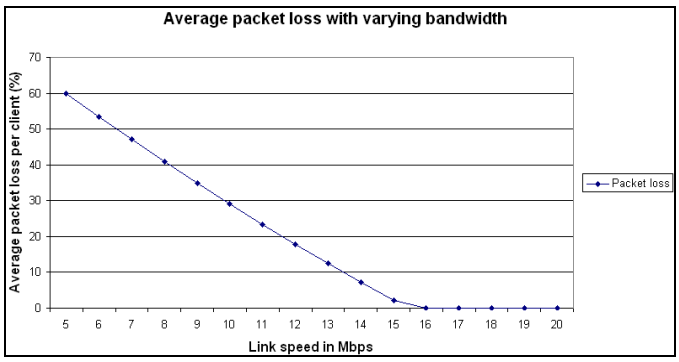


Figure 11. Average packet loss with increase in bandwidth

Bandwidth has a very profound effect on packet loss percentage, which in turn plays a major role in video quality. Factors like delay and jitter can be handled by the client with enough buffering. Packet loss on the other hand cannot be handled by the client without distortion or loss of video quality. With increasing bandwidth, packet loss is drastically reduced.

From the above analysis we can see that for 20 clients downloading a video of 1024Kbps at the same time, a minimum access network bandwidth of 16Mbps is required for optimal performance. The video on demand service provider is totally dependent on the access network which connects the VOD server to the client. Access networks can scale across multiple routers and may even cross multiple countries. It is thus very difficult on the part of the VOD service provider to ensure and guarantee video quality across the access network as the service provider has less to no control over the internet.

IV. CONTENT DELIVERY NETWORKS

The major factor that was found to affect video quality in centralized architecture is the access network link. This is totally removed out of picture when it comes to video delivery to the client in the CDN architecture. Figures 12 and 13 show an overview and the flow diagram of content delivery networks proxy video-on-Demand servers, denoted by red hexagons are placed close to the edge of the network. By placing the VoD proxies close to or at the ISP's infrastructure, the massive delays and losses incurred due to access network links can be avoided. The access network links are used only for replicating video content from the central VoD server to the VoD proxies.

In the below simulation, similar to centralized architecture, the clients are connected to the respective ISP's with different link speeds to simulate the current popularly available client bandwidth rates. Clients 1-5 were connected at the rates of 768 Kbps, 1024 Kbps, 2 Mbps, 4 Mbps, and 6 Mbps respectively. All clients were simulated to request a same video of length 750 frames. Various factors such as delay, inter-frame delay and jitter have been studied.

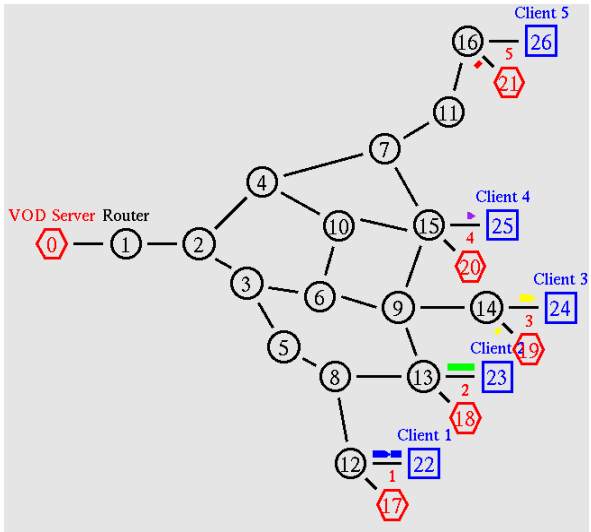


Figure 12. Simulated CDN (Hexagons labeled 1,2,3,4 & 5 are the proxy servers at the edge of the network)

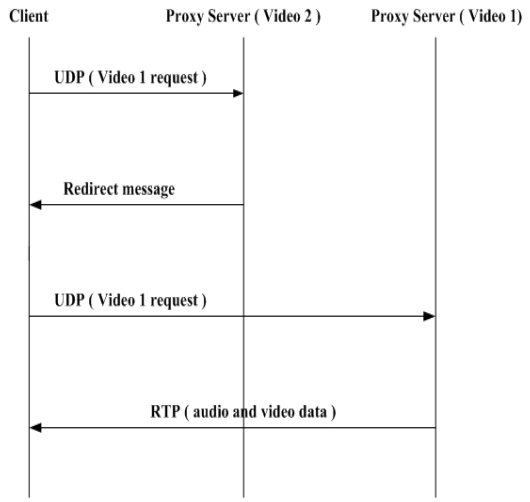


Figure 13. Flow diagram of Content Delivery Networks

In contrast to the centralized architecture, where the access network bandwidth and the client bandwidth both played a role in the delay noted, here only the client bandwidth affects the delay. This can be seen by comparison on data of clients 3, 4 and 5. In case of centralized architecture, due to better access link speed and delay client 3 connected at 2 Mbps was performing better than client 5 connected at 6 Mbps. In case of centralized architecture, the delay decreases with increasing client bandwidth rates.

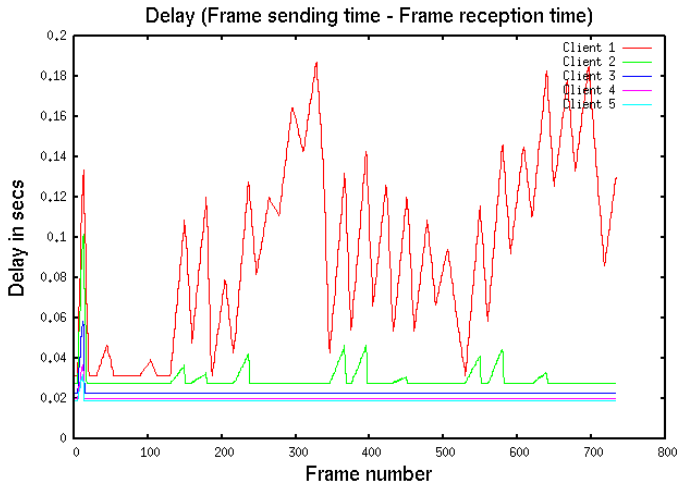


Figure 14. Delay for video transmission from proxy server to clients

The plots of Figure 14 clearly show the advantages of a distributed architecture as content delivery networks. Centralized architecture has a single point of failure and leads to heavy network congestion at the core of the network, which drastically affects the video quality. By distributing proxy servers towards the edge of the network, CDN no longer has a single point of failure. As the number of hops between the client and server are reduced - jitter, delay and packet loss are comparatively lesser and better video performance is achieved. Hence the deployment of proxy servers near to the clients makes the VoD systems more reliable and scalable.

V. CONCLUSION AND FUTURE WORK

In this article we analyzed the VoD traffic, and compared the performance of VoD architectures under different network conditions. We suggested efficient network parameters required for VoD architectures to become more reliable and scalable. With the best possible video on demand infrastructure, if the client does not make wise use of the existing bandwidth limits, video quality would be poor. Heavy network usage, computer virus, old versions of operating systems and unprotected network on the client end can have adverse effect on video on demand quality. Cable modems and wireless routers generally used at homes these days have good firewalling and quality of service configuration support. Bandwidth can be reserved in these devices for video on demand traffic. This ensures that even when in times of excessive network usage, minimum required bandwidth for video on demand traffic is available. The research in this article can be extended to non-centralized architecture in future.

REFERENCES

[1] T. Little and D. Venkatesh, “Prospects for Interactive Video-on-Demand”, IEEE Multimedia, Fall 1994, Volume: 1, Issue: 3, pp. 13-17.

[2] R. Louis, “Factors affecting video quality metrics,” Telchemy Publisher, February 2008.

[3] http://www.isi.edu/nsnam/ns/, Last access date Nov. 2010.

[4] C. Y. Yu, C. H. Ke, R. S. Chen, C. Shieh, B. Munir, and N. Chilamkurti, “MyEvalvid_RTP: a New Simulation Tool-set toward More Realistic Simulation,” Future Generation Communication and Networking, 6-8 Dec.2007, Volume: 1, pp. 90-93.

[5] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson,. “RTP-A Transport protocol for real-time applications,” July 2003, Request for Comments: 355.

 ISBN:978-1-61208-113-7

Usability Evaluation and Study of a Video-Conferencing Service Provided via the Virtual Conference Centre

Borka Jerman Blažič
Jožef Stefan Institute, Faculty of Economics, Ljubljana, Slovenia
jerman-blazic@ijs.si

Tanja Arh
Jožef Stefan Institute, Ljubljana, Slovenia
tanja@e5.ijs.si

***Abstract*—Usability evaluation is a core component of user-centred design (UCD) approach and aims primarily to evaluate effectiveness, efficiency and satisfaction when users interacting with a product/service to achieve their goals and needs, influencing their decision to its future adoption (i.e., user acceptance). The focus of this research work was to assess the usability of videoconferencing system with traditional usability method (usability testing) in order to get good assessment of its learnability and applicability. Based on the five tasks performed by 29 participants from five countries, the overall evaluation of the Virtual Conference Centre (VCC) was satisfactory and implicit a high level of reliability.**

Keywords - usability evaluation; task scenarios; videoconferencing system; collaborative environment.

I. INTRODUCTION

During the last years, sites like Google, Flickr, Youtube, LinkedIn, Facebook, Myspace, Skype and many more have developed extremely successful mass services which have led to a service paradigm known as Web 2.0. Web 2.0 is more an evolution in service design than a revolution. It proposes the use of a set of basic principles such as: "Web as platform", putting the user in the centre (rather than the technology); it uses simple user interfaces, aggregate knowledge and wisdom of crowds by using folksonomies, uses social software, makes proper and extensive use of URIs (Uniform Resource Identifiers) and HTTP using REST (Representational State Transfer) and Web applications, uses peer-to-peer and Grid networks, etc.

On the other hand, the use of videoconferencing and collaboration tools in the Internet is slowly taking up due to the availability of more bandwidth and to the development of better integrated and more usable tools. Those tools do not inter-work in most cases and are used mainly today for realising relatively simple tasks like working meetings, substituting the popular conference calls, to connect remote speakers on co-located conferences or meetings, etc. The integration of videoconferencing and collaborative tools within a Web 2.0 service is leading to gather use of those tools.

The growth in videoconferencing and collaboration tools and systems is heralding also a shift in the nature of Human Computer Interaction [1]. Use of new media technologies such as videoconferencing systems is becoming a part of everyday communication and entertainment amongst individuals. Indeed, user satisfaction with technologies related to distance and collaborative applications is an integral part of usability [6], which is defined as the extent to which people can use the product *quickly* and *easily* to accomplish their tasks [2]. Usability testing is to make sure that users can find and work with the functions of the product to meet their needs. One of the most important outcomes of usability test is a list of problems, which entail changes and thus improvement of the product. Usability evaluation of videoconferencing and collaboration tools is traditionally conducted by means of task performance measures and subjective measures such as questionnaires, interviews, etc.

The paper is organized as follows: the background and production description section introduces the application studied; the evaluation methodology and participants section describes the approach used. The next section of the paper deals with the task scenarios and the last section discuss the results. The paper ends with short conclusion.

II. II. BACKGROUND AND PRODUCTION DESCRIPTION

The Virtual Conference Centre (VCC) was designed following the Model-View-Controller Model together with agile software design methods such that the logic and the models of the various functionalities can be developed independently of the views. The software design frameworks *Ruby on Rails and RESTful* were chosen, as they are the most efficient rapid prototyping tools for website design.

 ISBN:978-1-61208-113-7

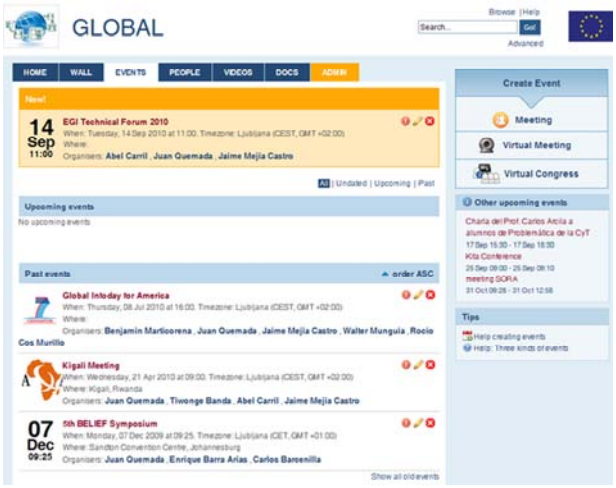


Figure 1: Event Space in the VCC

The VCC includes the following main features:

- A single unified point of access to the virtual auditorium features enabling access to the different functionalities through the available tabs: *Home, Events, Posts, People,* and *Spaces.* It also contains a specific button for direct access to the GLOBAL project, another for description of the project partners, a *Login* and a *Registration* button.
- *Spaces* are the means to organise different projects and different topics. Each space has a repository with public and private documentation and events (this functionality was not included in the usability testing). The space's public area can be customised to be the public face of the project in the VCC.
- In the spirit of openness, any user may register to the VCC. The registration procedure includes spam protection. The VCC registration allows the user to access publicly available spaces, documents, event announcements, etc. More importantly, it allows granting access to areas of the VCC that are marked *private* by a particular group of people.
- A profile display for every registered user such that he/she can update the information stored, such as password, address, email contact, project interests, etc. Profiles will facilitate user networking and partnership building.
- The *People* tab displays the registered users for quickly finding specific contacts and can only be seen by registered users.
- The *Spaces* tab shows the available VCC spaces. Each *Space* has a space administrator. The space administrator can add users to or delete them from a specific space, allow an "open to all" registration policies, creates user groups under that space, and even delete the complete space.
- The *Events* tab is the event scheduler and the main part of the VCC. It allows the registered user to create events and organise event documentation in a clear accessible manner assigning access privileges to each item. Any VCC visitor, even unregistered users, may see the event calendar with all the public events, incentivising new registrations. Only registered users can access private events of the spaces he is a member of.
- The *Posts* tab gives access to the repository of shared documents and a blog-like or forum-like upload system based on posts to fill it. These posts may have multiple attachments. In future versions of the VCC other upload mechanisms will interface the user with the repository.

Virtual Conference Centre has been developed by the GLOBAL project from EU FP7 program.

III. III. EVALUATION METHODOLOGY

Standard user test procedures were adopted [5]. Jožef Stefan Institute (JSI) playing the role of *General UT Coordinator* was responsible for the overall coordination of UT, the compilation and analysis of the data, and the production of a usability report to be sent to the development team (UPM). The key role of *Local UT Coordinator* was to coordinate the conduction of the usability tests in his or her site and to ensure the required data would be collected and sent to General UT Coordinator. To ensure the uniformity of the testing procedures and thus the validity and reliability of the testing results, a document entitled "*GLOBAL Usability Test – Questionnaires & Task Scenarios*" describing in detail the goals, instruments, participant requirements, procedures, data handling, etc. Local UT Coordinators were supposed to follow the Guidelines with minimum deviations. The language of communication between General UT Coordinator, Local UT Coordinators and Local Testers is written English. Hence, it is expected that all of them possess certain level of English proficiency.

IV. PARTICIPANTS

The profile of the participants was characterized by several elements:

A. General characterization of real user population

Companies, researching and administrative staff of research institutes and higher education institutes, including director, professor, lecturer, tutor, and researcher, who have certain level of experiences and knowledge with regard to ICT can be users of the VCC. Characteristics which were common to all test participants were:

 ISBN:978-1-61208-113-7

- Possessing experiences of using software applications;
- Possessing some basic knowledge of ICT and about videoconferencing systems, etc.;
- Possessing good English proficiency, at least a high level of reading comprehension.

B. *Characteristics of the testing sites*

English language version of the VCC has been tested in UbuntuNet in Malawi, Zentrum für Soziale Innovation in Austria, Jožef Stefan Institute in Slovenia, U. Politécnica de Madrid (UPM) in Spain, University College London in United Kingdom and CLARA in Peru.

C. *Characteristics of the testing participants*

The number of participants involved in the usability tests was 29 (8 female and 21 male). All of them had the educational level of at least the first university degree. Their participations were voluntary. Prior to working out the task scenarios with the VCC, the participants were required to complete a Pre-test Questionnaire on demographic data (gender, age, job title). This questionnaire also reflects the average level of competence in ICT (M = 3.80) and the average level of competences in videoconference systems (M=3.09). We convert the nominal to interval scale, with left anchor "1" indicating the lowest level and right anchor "5" the highest level of the attribute in question. None of the participants had interacted with the VCC before they took part in the usability tests. These demographic data are relevant for interpreting the results of usability tests.

D. *Profile of the local tester*

Ideally, Local Testers should be usability specialists or highly experienced in conducting usability tests. However, it may not be easy to get them, especially when the resource is tight. Alternatively, those who meet the following criteria can assume the role of Local Tester:

- Experienced in conducting experiments with human participants,
- Has some knowledge in Human Computer Interaction (HCI),
- Fluent in the native language and English,
- Motivated to do research,
- Good at observation,
- Able to manage several tasks at one time.

It was recommended to employ the same Local Tester to conduct all the testing sessions to ensure the consistency of data recording and interpretation.

V. TASKS SCEANRIOS AND PROCEDURE

A set of five tasks covering the core functionalities of the VCC was employed. The tasks were presented in English. All test participants possess a reasonable level of English reading ability. Below is the list of the tasks:

- *(T1) Obtaining a user account for a Virtual Conference Centre (VCC)*
- *(T2) Creation of new space in VCC and joining an existing one*
- *(T3) Creation of new event*
- *(T4) Modifying the event in event manager*
- *(T5) Sending private message*

Each of the above five tasks was translated into task scenarios, which render the test more realistic and problem oriented (e.g., *You are organising a workshop or other distributed event like a meeting. Therefore, you need to create an event in the space you created in task 2. Log yourself into the VCC and create a new event. The event must be marked as Isabel event).* In addition, for each of the task scenarios, quantitative usability goals in terms of task completion time and number of errors were set, which were benchmarked by an experienced user of the VCC. They can serve as references or baselines for data analysis. System Usability Scale (SUS) (Brook, 1996) and Feedback Questionnaire (FQ) (After-Scenario Questionnaire) [4] were employed. Test participants were welcomed and briefed about the goal and procedure of the usability tests, which was followed by an explanation of the equipment to be used. Participants were asked to perform a set of selected task scenarios that cover most frequent as well as critical functionalities of the VCC.

After each task, participants were asked to complete the After-task questionnaire, consisting of four questions (Q1-Q4), which were derived from the literature on usability research [4]. A 7-point Likert scale was employed with left anchor indicating lowest level of satisfaction and right anchor the highest. Q.3 and Q.4 evaluated the same two variables, which are nonetheless phrased. After completing all the five tasks, participants were asked to complete Post-test questionnaire entitled "System Usability Scale (SUS)" which consists of 10 questions and has psychometric properties.

VI. ANALYSIS AND RESULTS

We have categorized collected data along two dimensions: (i) qualitative vs. quantitative and (ii) objective vs. subjective (Table 1). Some analysis results of these data types are presented in subsequent sections.

TABLE 1: TWO DIMENSIONS OF DATA TYPES

	Qualitative	Quantitative
Objective	A list of usability problems (UPs) derived from the participants' notes	Time-on-Task Effectiveness & Efficiency
Subjective	Participants' comments in SUS	Responses to the questionnaires: ▪ Pre-test Q. ▪ Feedback Q. ▪ Follow-up Q. (SUS)

Descriptive statistics – mean and range – comprehends six quantitative objective and subjective usability measures: duration (min), perceived ease of use,

 ISBN:978-1-61208-113-7

perceived efficiency, perceived difficulty, perceived time-consumingness and task completion rate (%).

A. *Quantitative data*

Different quantitative measures were taken, including time-on-task, proposed time, mean time and standard deviation. The two usability metrics – **effectiveness** and **efficiency** – were derived effectiveness and efficiency.

Furthermore, effectiveness and efficiencies per task were computed. Effectiveness denotes the rate that a task is completed successfully *without* assist from any help desk – unassisted completion rate. Efficiency is calculated through dividing an unassisted completion rate by its corresponding unassisted mean time-on-task. In the Table 2, for the sake of comparison, both unassisted and assisted completion rates together with their corresponding mean time-on-task are displayed. The average effectiveness over five tasks was 88.57 %. All participants could complete all the given tasks, with or without assist from the Local Tester. The average efficiency over five tasks is 21.52 %/minute, ranging from 7.14 %/min (Task 4) to 33.33 %/min (Task 1). In fact, Task 4 (Modifying the event in event manager) was proved to be problematic.

B. *Time on tasks*

Each participant was required to perform five tasks. Based on the data of the 29 participants, the value (average time) of this variable is 30.65 minutes, with the range from 15.00 (JSI-P5) to 81.00.

TABLE 2: EFFECTIVENESS & EFFICIENCIES PER TASK

Task	Total Completion (with or without assist)		Effectiveness (%) (tasks without assist)		Efficiency (%/min)	Total no. assists
	Rate (%)	Mean Time	Rate (%)	Mean Time		
1	100 %	4,72	100.00%	3,00	33.33	0
2	100 %	4,79	100.00%	4,00	25.00	0
3	100 %	6,20	85.71 %	5,00	17.14	1
4	100 %	9,48	57,14 %	8,00	7.14	3
5	100 %	5,44	100.00%	4,00	25.00	0

Altogether, 145 tasks were performed and 145 (100 %) were successfully completed. As shown in Table 2, among the 5 tasks, Task 4 (Modifying the event in event manager) was found to be most problematic. Indeed, the average time-on-task for Task 4 is 9.48 minutes, exceeding the benchmarked upper bound (i.e., 6.00 minutes) by 63.2 %. The range of time-on-task for Task 4 is large, spanning from 4.00 min (UPM-P2) to 20.00 min (UBN-P1 and UBN-P6). As a matter of fact, Task 3 (Creation of new event) is quite similar to Task 4 (Modifying the event in event manager) except the additional scheduling. As evidenced by the data, the average time-on-task for Task 3 is less than that for Task 4. All participants performed Task 3 much faster. However, in some other cases, the reverse could be observed (e.g., UCL-P2). This may be attributed to the fact that this user found the task very confusing and without logic.

One of the most important results from the Usability Tests is the list of Usability Problems (UP) identified when the participants interacted with the system to achieve the given tasks (Table 3). The six testing sites have collected different sets of UPs, with a number of them being overlapped and the others being unique. We compiled and integrated the six lists of the usability problems into a complete list (see Table 3). The implications of individual columns are:

- *Usability problem (UP)*: It is the identifier of individual UP.
- *Task ID*: It denotes in which task the UP was identified. For instance, Task ID 2(5) means that this UP was identified in Task 2 (Creation of new space in VCC and joining an existing one) and number 5 in brackets indicating how many of the test participants found the problem.
- *Descriptions of Usability Problem*: Detailed explanations what the UP was and how the UP was identified.
- *Severity:* There are three levels:
  - *Moderate usability problems* are those that significantly hinder task completion but for which the user can find a work-around.
  - *Severe usability problems* are those that prevent the user from completing a task or result in catastrophic loss of data or time.
  - *Minor usability problems* are those that are irritating to the user but do not significantly hinder task completion.

As shown in Table 3, there are altogether 21 the most important usability problems (UP). Some of the UPs have frequency only once and the highest frequency is 12 for UP13 (Task 3: Create a new event and invite people to this event) - users had difficulty in inviting people to the event, as well as the option to invite people during the event creation. In a scenario-based usability study, participants use a computer application to perform a series of realistic tasks. The FQ is a 3-item questionnaire to assess participant satisfaction after the completion of each scenario [3]. The items address three important aspects of user satisfaction with system usability: *ease of task completion, time to complete a task,* and *adequacy of support information (online help, messages, and documentation*). Each item is rated with a 7-point Likert scale, with 1 being "Strongly disagree" and 7 "Strongly agree". The items are phrased in a positive manner. Hence, the *higher* the score, the more the user is satisfied with the system. The questionnaire takes very little time for participants to complete. Table 4 shows the results of FQ of the five tasks. Q1.1 addresses the ease of task completion for Task 1 as perceived by a user; Q1.2 addresses the degree to which the user is satisfied with the time to complete Task 1.

 ISBN:978-1-61208-113-7

TABLE 3 IMPORTANT VCC USABILITY PROBLEMS

UP	Task ID	Descriptions of Usability Problem	Severity
UP1	1(9)	The participant hardly found the »Login« button on the start page.	Moderate
UP2	1(5)	The Login/Register button at the top of the page is quite inconsistent. Clicking the button you get the login page.	Moderate
UP3	1(3)	No alternative way for registering or troubleshooting while registering: contact info who to turn to if something goes wrong (e.g. not receiving confirmation mail).	Minor
UP4	1(2)	The facility for password reminder transmission via email did not work. The request was submitted successfully, a confirmation was displayed, but no email with the password was received.	Severe
UP5	2(2)	Topic titles/names are presented very user unfriendly (e.g. spaces/spacesname/post/200).	Minor
UP6	2(1)	User spaces should be allowed to be created only by registrated users (those who has confirmed registration).	Minor
UP7	2(9)	The purpose of groups within spaces should be more clearly explained.	Severe
UP8	2(2)	No connection to other users of VCC from your space (invitations).	Minor
UP9	4(7)	Confusion between "Comment" and "Post".	Severe
UP10	4(4)	Adding attachments to posts/comments is a little confusing – only one item can be added – no confirmation is given.	Minor
UP11	4(7)	When displaying "The post can't be empty" for e.g (or other system info) there's no exact explanation on the specific error made in the post.	Moderate
UP12	General (2)	Sometimes operations are performed without giving feedback/notice to user.	Moderate
UP13	3(12)	The "invite people" option is not intuitive to find, as well as the option to invite people during the event creation.	Moderate
UP14	3(9)	Private Message to complex to sent.	Moderate
UP15	General (6)	Help is not available.	Moderate.
UP16	General (3)	No information about what VCC is what it offers, what are its features.	Minor
UP17	General (3)	The participant hardly found the link to the support & help.	Minor
UP18	General (5)	Look of the VCC; everything happens in the most right part of the screen.	Minor
UP19	Invitation	The system crashed without any warning or error messages	Severe
UP20	General (3)	The Admin tab should be differentiated from the other tabs (e.g. Home, Posts, Events, etc).	Minor
UP21	General (5)	Clicked the back button of the browser and lost all the data, very frustrating.	Severe

TABLE 4. RESULST OF THE FQ TASKS

Qn	Q1.1	Q1.2	Q1.3	Q2.1	Q2.2	Q2.3	Q3.1	Q3.2	Q3.3	Q4.1	Q4.2	Q4.3	Q5.1	Q5.2	Q5.3
Mean	5,41	5,66	5,07	5,90	5,90	5,59	5,21	5,38	4,72	4,38	4,48	3,76	4,48	4,79	4,07
St. Dev	1,72	1,72	1,44	1,08	1,35	1,38	1,29	1,40	1,36	1,68	1,60	1,62	2,03	2,38	1,96
Min	2,00	2,00	2,00	3,00	2,00	1,00	2,00	2,00	2,00	1,00	2,00	1,00	1,00	1,00	1,00
Max	7,00	7,00	7,00	7,00	7,00	7,00	7,00	7,00	7,00	7,00	7,00	6,00	7,00	7,00	7,00

Q1.3 addresses the adequacy of support information for Task 1 as perceived by a user. The same sequence is for Task 2 to Task 5. For Task 2 (Creation of a new space in the VCC), the ease of completion was rated as 5.90, the degree of satisfaction with the completion time was 5.90 and the adequacy of support information was 5.59. These ratings imply that the users generally were quite satisfied with this particular Task. Task 1 (Obtaining a user account in the VCC) has similar ratings, but of lesser degree. Task 4 (Modify and event in event manager) and task 5 (Send private message) imply that the users generally were not so satisfied with this particular tasks.

Based on the five tasks performed by 29 participants with different cultural and academic backgrounds as well as various levels of experiences and knowledge in information technologies and video-conferencing systems, the overall evaluation of the design of the VCC (beta version) was satisfactory. The English language version of the VCC has been tested independently in six different sites. The average effectiveness and average efficiency of the five tasks over the 29 participants are 88.57 %.

VII. CONCLUSION

The results of the study show that the users' performance is highly acceptable to be improved to render it suitable for a wider scope of users, especially those who have limited experience and competence in ICT and in the domain of video-conferencing systems.

ACKNOWLEDGMENT

This work has been performed in the framework of the EU funded FP7 project GLOBAL.

REFRENCES

[1]. T. Brinc, D. Gergle and S. D. Wood, Usability for the Web: Designing Web Sites that Work. San Francisco, USA: Morgan Kaufmann, 2002.

[2]. J. Brooke, SUS: A 'quick and dirty' usability scale. In W. Jordan, B. Thomas, B.A. Weerdmeester, I. L. McCleland (Ed.), Usability evaluation in industry (21, pp. 189-192). London, UK: Taylor & Francis, 1996.

[3]. J. R. Lewis, Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. SIGCHI Bulletin, 23(1), 1991, pp.78-81.

[4]. J. R Lewis, IBM Computer Usability Satisfaction Questionnaires:Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 7(1), 1995, pp. 57-78.

[5]. J. Nielsen, International usability testing. In E. del Caldo & J. Nielsen (Eds.), International user interface. New York: John Wiley & Sons, 1996.

[6]. A. S Patrick, The Human Factors of Mbone Videoconferences: Recommendations for Improving Sessions and Software. Ottawa. CRC Technical report, 2006.

 ISBN:978-1-61208-113-7

Multi-Episodic Dependability Assessments for Large-Scale Networks

Andrew P. Snow and Andrew Yachuan Chen
Ohio University
School of Information and Telecommunication Systems
Athens, Ohio
e-mail: asnow@ohio.edu

Gary R. Weckman
Ohio University
Department of Industrial and Systems Engineering
Athens, Ohio
e-mail: weckmang@ohio.edu

Abstract— **As a network infrastructure expands in size, the number of concurrent outages can be expected to grow in frequency. The purpose of this research is to investigate through simulation the characteristics of concurrent network outages and how they impact network operators' perspective of network dependability. The dependability investigated includes network reliability, availability, maintainability and survivability. To assess this phenomenon, a new event definition, called an "impact epoch", is introduced. Epochs are defined to be either single, concurrent, or overlapping outages in time, which can be best assessed with new metrics and simulation. These metrics, Mean-Time-To-Epoch, Mean-Time-to Restore-Epoch along with percentage time the network is not in an epoch state (Quiescent Availability) and Peak Customers Impacted, are investigated. A case study based upon a variable size wireless network is studied to see what insights can be garnered through simulation. The new proposed metrics offer network operators valuable insights into the management of restoration resources. Simulation proved invaluable in identifying multi-outage epochs, as modeling their occurrence, frequency, duration and size is analytically intractable for large networks.**

Keywords-simulation, survivability, reliability, maintainability, wireless network infrastructure

I. INTRODUCTION

Telecommunication networks have become critical telecommunication infrastructure as millions of people depend on these networks for daily communication and commerce. As demand increases, so does network size, challenging engineers and operators to maintain and not compromise network dependability. As a network grows in size, the sheer number of components grows also, increasing failure hazard. With such an increase in hazard, the chance of concurrent, or overlapping, outages also can be expected to increase. Dealing with these concurrent outages is challenging because network operators have to judge priorities in allocating limited repair resources to outages spatially distributed. If the response is consistently substandard, the operator's ability to satisfy current and accommodate new customers could be adversely affected. Understanding the characteristics of concurrent outages as a function of network size and component failure and repair rates offers network operators valuable information in developing outage recovery strategies. The number of customers that could be impacted by network failures is another important factor for network operators to consider. If the probability distribution of impacted customer is known, thresholds highlighting critical events can be established. This paper investigates the characteristics of simultaneous network outages and attempts to identify the distribution of impacted customers through simulation.

A. *Dependability*

Dependability has a number of different attributes. According to Laprie [9], the concept of dependability includes attributes like availability, reliability, maintainability, safety, confidentiality and integrity. Others have included survivability as an additional network dependability attribute, since it is so important to measure the resiliency of the network to provide partial service to the population of users during network service disruptions [8]. The higher the survivability, the better chance a service provider has to satisfy customers in times of network stress due to component failures or traffic overloads. Integrity and confidentiality are not considered in the scope of this study. Rather, we consider ARMS attributes (availability, reliability, maintainability, and survivability) of dependability.

B. *Reliability*

Network reliability is defined as the probability that a network will perform its required functions over a specific period of time [10]. The reliability, for a network or a network component is expressed as the probability that a network or component will not fail over some specified time period of interest, given by [11]:

$$R(t) = e^{-\lambda t} = e^{-t/MTTF} \quad (1)$$

Where λ is expected failure rate and MTTF is the expected average time between failures. If the time-period of interest is reasonably short, MTTF is assumed to be constant, meaning that an assumption of a Homogeneous Poisson Process (HPP) can be made.

C. *Maintainability*

Network maintainability is defined as the ability of a network to recover from failures [11]. Maintainability can be determined from the Mean Time to Restore (MTR). Restore time is a random variable and typically consists of three parts – detection time, travel time to the outage

 ISBN:978-1-61208-113-7

location and the actual repair or replacement time. In this research, the lognormal distribution is used since travel time plays an important role.

D. Availability

Network availability is defined as the probability that a network is ready for use when needed [11]. Average availability can be expressed as:

$$A = \frac{MTTF}{MTTF + MTR} \quad (2)$$

Availability is a good metric to assess the state when the network is experiencing no problems due to failures.

E. Survivability

Network survivability is defined as the ability of a network to provide services to most customers under partial failures. Snow [13] defined Prime Lost Line Hour (PLLH) as an impact measure for wire-line network outages that take into consideration usage levels at the time of the outage. PLLH is the product of the estimated number of customers impacted and the duration of an outage. Total Line Hours (TLH) is the product of the total number of customers served by the network and the total hours in the time-period of interest, resulting in a network survivability calculation in Equation (3).

$$NS = 1 - \frac{PLLH}{TLLH} \quad (3)$$

The Telecommunication Committee T1, an ANSI certified standards organization, developed the "outage index" as a survivability metric that includes consideration of the size and duration of the outage, in addition to the importance of the services affected by the outage. This metric uses weights for each of these three dimensions, and has been shown to be a questionable metric [15, 14, and 4].

II. IMPACT EPOCH

The focus of this research is on concurrent and time-overlapping component outages as the network size scales. In order to describe the characteristics of concurrent or overlapping outages from a network operator perspective, a new concept called *impact epoch* is introduced. An impact epoch starts when a network transfers from the state of no customers impacted to a state of having customers impacted; and it continues until the network returns to the state of having no customers impacted. An impact epoch event includes single or multiple outages that overlap in time. The number of impacted customers during one impact epoch is not necessarily constant, since a single impact epoch may include more than one component outage due to nearly simultaneous failures in the network. An example of a single impact epoch, which consists of three overlapping outages, is shown in the Figure 1 in the form of an epoch profile. Time is represented by the X-axis and the Y-axis represents the percentage of customers that can be served in the network. Prior published research has not considered an epoch perspective; hence, this new methodology is investigated in this paper.

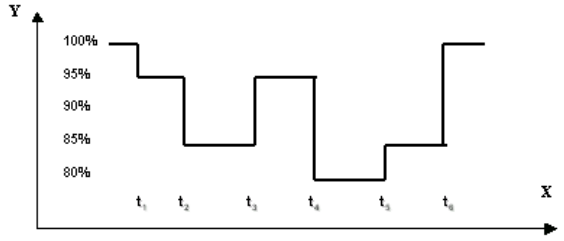


Fig. 1. Single impact epoch due to three overlapping outages

Since epochs are arrival events, MTTE is defined as the mean time to impact epoch in a network. MTTE offers insights into the average interval before operators can expect disturbances that render the network incapable of satisfying all customers. Longer MTTE implies that the network has higher reliability, or the capacity and performance to lesson congestion events. Since epochs have duration, MTRE specifies the mean impact epoch restore time - a description of a network's maintenance response, or ability to gracefully recover from congestion. Shorter MTRE implies that the network has better maintainability or recoverability. MTRE together with MTTE provides the average quiescent time (A_Q), or the fraction of time the network, on average, is not undergoing a disturbance that impacts customers. Quiescent availability can be determined by the following equation:

$$A_Q = \frac{MTTE}{MTTE + MTRE} \quad (4)$$

Survivability from an epoch perspective can still be measured by Equation 3. However, in an environment where there may be concurrent or overlapping outages, peak customers impacted (PCI) may be of interest. For instance, in Figure 1, the PCI is 20%.

The advantages of studying impact epochs instead of a single outage are that epochs:

- Provide a better-detailed description of the cumulative time phased effect of network disturbances
- Offer a new way to evaluate network dependability, providing a different perspective important to network operators
- Provide insights into how characteristics such as frequency, duration, number of concurrent outages, and peak customers impacted might change as network size varies

Table 1 illustrates the mapping between wireless network dependability attributes and the metrics developed in this paper to assess them. In this wireless network example, a Wireless Traffic Profile (WTP) is developed using empirical wireless traffic data from the literature, allowing computation of PCI and WPLLH (Wireless prime lost line hours). In this study, outages are due to component failures. In other words, this is a fault management rather than a performance management perspective -- operators are

 ISBN:978-1-61208-113-7

responding to outage events induced by component failures, and the need to restore or replace the faulty components. Therefore, this work presents conservative estimates of episodic occurrences.

TABLE 1

New Network Dependability Metrics

Dependability	Network Attribute Name
Reliability	Network Mean Time To Epoch (MTTE)
Maintainability	Network Mean Time Restore Time (MTRE)
Availability	Network Quiescent Availability (A_Q)
Survivability	Peak Customer Impacted (PCI)

III. WIRELESS NETWORKS

Extensive research has been conducted over many years regarding the traditional wire-line telephone network, also called the Public Switched Telephone Network (PSTN). These research efforts helped wire-line networks offer very dependable services with a common quality metric of Five 9's availability [3]. On the other hand, research in the world of wireless communication, especially in cell phone networks, is by comparison relatively new. Research into wireless telephone network reliability did not receive much attention until the late 1990s. Over the last 15 years, the wireless network has grown at an amazing rate. According to the Cellular Telecommunications Industry Association (CTIA) wireless Quick Fact Sheet [18], cellular subscribers in the US surpassed 5 million in 1990 and doubled in just two years. By 2000, cellular subscribers exceeded 100 million in the US and wireless penetration rate was over 65%. There were over 182 million customers in the US as of May 11, 2005.

In 1992, the FCC at first ruled that wire-line carriers had to report all outages that impacted more than 50,000 customers for at least 30 minutes. This threshold was quickly lowered to 30,000 customers for 30 minutes in 1993 [12]. Statistical failure data of wire-line local switches are publicly available from the FCC's Automatic Reporting and Management Information System (ARMIS) database. However, starting January 2, 2005, the FCC ruled that wireless carriers also had to report their network outages to the FCC [6]. Meanwhile, the FCC established a four-year rollout plan for E911 phase II, which began in October 2001. Phase II required wireless carriers to provide precise location information for wireless 911 calls, within 50 to 300 meters in most cases [7].

A. *Wireless Network Infrastructure*

The general structure of a wireless network with most of the required functional components is shown in Figure 2 [5]. They include the network operation subsystem, base station subsystem and network switching subsystem. Each subsystem includes a number of components that are studied in this research. This is a 2G+ architecture that has some similarity to 3G/4G architectures from hierarchical and topological perspectives. The Base Station Subsystem (BSS) is comprised of Base Stations (BS) and Base Station Controllers (BSC). A BS is essentially the radio station that broadcasts to and receives from the mobile station in a "cell". A BSC is the controlling node for one or more cells or BSs and manages voice or data traffic and signaling messages for all the cells under its control. The BSS provides the transmission path including traffic and signaling between mobiles and the Network Service Subsystem (NSS) [5].

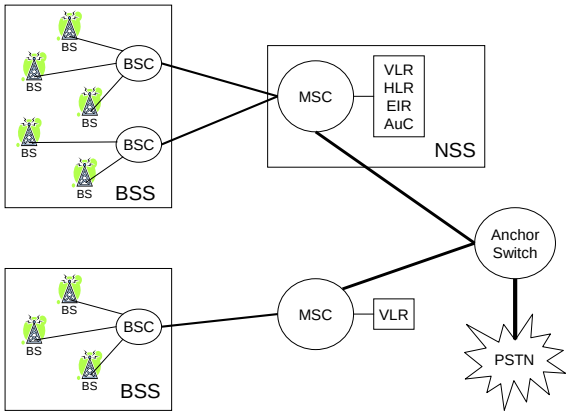


Fig. 2. Wireless network infrastructure

The NSS is the switching and control portion of the entire wireless network. It is comprised of the Mobile Switching Center (MSC) and three intelligent network nodes known as the Home Location Register (HLR), Visitor Location Register (VLR), Equipment Identity Register (EIR), and the Authentication Center (AuC) [5]. The MSC is the central heart of a wireless network. The failure of a MSC typically results in communication loss of all users that the MSC controls, since calls cannot be originated or terminated. Carriers pay close attention to the status of a MSC since it supports billing functions such as collecting Call Detail Records (CDR). A typical MSC is engineered to be highly reliable. In A. Snow, [16], the authors introduced a wireless network infrastructure called the Wireless Infrastructure Block (WIB). The scope of the WIB is from the BS to the MSC including the HLR/VLR database. They also discussed how MTTF and MTR in a WIB might affect the network's dependability [16]. The topology used in a WIB is the star topology. Large wireless infrastructures consist of multiple WIBs.

B. *Wireless Traffic*

Advantages of using the star topology include supporting modular expansion, and simplified monitoring and trouble-shooting. The largest disadvantage of star topology is the

 ISBN:978-1-61208-113-7

creation of single point of failure, such as the MSC and database. Fortunately, these components are highly reliable. Table 2 indicates the number of components in a WIB along with the number of customers potentially impacted by each component. A WIB can serve up to 100,000 customers. How many subscribers are actually impacted depends on utilization, which can be related historically to time of day and day of week. This can be represented by a time factor, which is really a time phased traffic profile that reflects percentage utilization at a point in time [17]. According to historical statistics [13], heavy traffic load in wire-line networks occur between 9:00am and 4:00pm on weekdays.

TABLE 2

Number of Components in One WIB and Maximum Failure Impact

Component	Number in One WIB	**No. Customers Potentially Impacted**
MSC	1	100,000
VLR/HLR DB	1	100,000
MSC-BSC link	5	20,000
BSC	5	20,000
BSC-BS link	50	2,000
BS	50	2000
Anchor-MSC Link	N	100,000
Anchor Switch	N	N * 100,000
Anchor Link	N	N * 100,000

Note: N is the number of WIBs in the wireless infrastructure

In this work, a new traffic profile for wireless networks is developed. The reason is that traffic patterns in wireless networks are different from that in PSTN. For instance, service charges in the PSTN are usually a flat monthly charge, while in a wireless networks there are more usage plans with differential charges based on time of day a call is placed. For example, many cell phone plans offer free calls at weekends and after 9:00pm on weekdays. Some people could wait until 9:00pm to place calls and take the advantage of this plan. Such phenomena results in different weekday and weekend traffic profiles in wireless networks. In Albaghdadi and Razvi [1], the authors studied an actual 1320 cell GSM network. In this research, the results reported in this GSM network were used to develop five-day weekday traffic and weekend traffic profiles as shown in Figures 3**.** These profiles were developed to create a wireless PLLH outage impact metric, called hereafter the WPLLH.

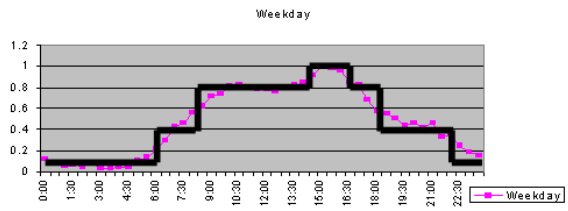


Fig. 3. Wireless Weekday Time Factor

Because the interaction of reliability and maintainability attributes are expected to be complex when it comes to investigating multi-episodic events, three different scenarios are investigated as follows: nominal, degraded maintainability, and enhanced reliability and maintainability. The nominal scenario signifies that the network is operating within published reliability and maintainability norms where regular maintenance schemes are used and reliability is stable. The degraded maintainability implies that the maintainability of the network is not as good as nominal, which signifies higher restore times from component failures. The enhanced reliability/maintainability scenario indicates that component reliability and maintainability are improved over nominal (with higher MTTFs and lower MTRs).

C. Network Component MTTF and MTR

Transmission links can be deployed with protection channels, wherein if the primary link is disrupted, the system switches to a protection channel**.** The more customers affected, the more likely there is a protection channel. Table 3 details a complete list of component MTTFs used in this study.

TABLE 3

Component MTTF and MTRs Used In the Study

Component Name	**Nominal MTTF (Yrs)**	**Enhanced MTTF (Yrs)**	**Degraded MTR (Hours)**	**Nominal MTR (Hours)**	**Enhanced MTR (Hours)**
Anchor Link	8.0	8.0	12.0	4.00	2.00
MSC-Anchor Link	8.0	8.0	12.0	4.00	2.00
MSC-BSC Link	2.7	4.0	12.0	6.00	3.00
BSC-BS Link	1.7	2.7	12.0	6.00	3.00
MSC and anchor switch	7.5	7.5	0.51	0.17	0.12
VLR/HLR database	3.0	4.5	2.00	1.00	0.50
BSC	3.0	6.0	4.00	2.00	1.00
BS	2.0	4.0	4.00	2.00	1.00

The nominal MTTF for other components was taken from [16]. As the MSC has become a very stable control and switch system over many years' development and deployment, in this case, the nominal MTTF and enhanced MTTF of MSC are taken to be the same, which is 7.5 years based on the results derived from empirical local switch statistics in the Federal Communication Commission's ARMIS database.

A component's maintainability is represented by its MTR. In order to understand the role that MTR plays in dependability, three MTR scenarios are used in the simulation: nominal, degraded and enhanced. Nominal MTR was obtained from [16]. The degraded MTR was taken as three times the nominal MTRs except for switches. Table

3 also lists the component MTRs used. The repair distributions are modeled based on a lognormal distribution, which is commonly used for long tailed distributions when travel time is involved. To summarize:

- The nominal case uses reliability and maintainability levels from literature and empirical data
- The enhanced case uses improved reliability and maintenance levels
- The degraded case uses lower maintainability levels

D. Simulation Model

Inputs for the simulation include all component MTRs and MTTFs, wireless traffic profile, the network size and an operational time of one year. Outputs from the program are network survivability, detailed outage information including start time, stop time, the number of customers impacted and the WPLLH for each outage. Other results like MTTE, MTRE, PCI and quiescent availability are derived from these simulation outputs using MS Excel™. Figure 4 displays the input and output process of the simulation and the derived results.

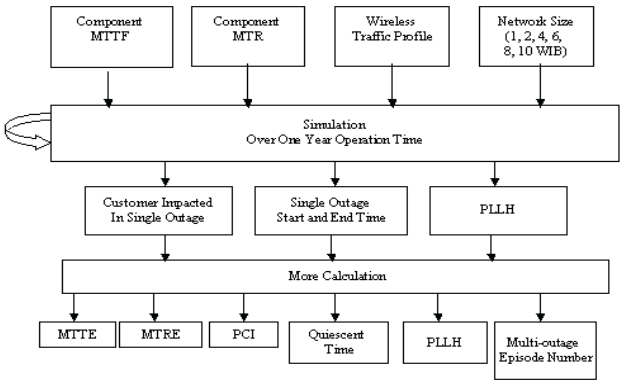


Fig. 4. Process of simulation and results

The process was conducted separately for different size networks (size determined by the number of WIBs) based on three scenarios: nominal, degraded maintainability, and enhanced reliability and maintainability. The maximum deviation in the nominal scenario between the simulation output and the analytical result was 0.85% for 8 WIB's, which was acceptable. This verified the simulation. Direct simulation program outputs include outage numbers, start time, end time, impacted customers, WPLLH and duration of each component outage. An example of a simulation output is revealed in Table 4, showing four component outages, starting at 308.465 days into the year. Figure 5 illustrates the impact epoch over the simulation time. The Quiescent Time can be derived from direct outputs of the simulation program and is calculated as:

$$Q_t = \sum_{i=1}^{n} TTE_i = TotalSimulationTime - \sum_{i=1}^{n} TRE_i \qquad (5)$$

where n is the number of quiescent periods. The sum of all TTEs and all TREs should equal the total simulation time, as shown in Figure 5.

TABLE 4

Simulation Output Example for A 10 WIB Network

Failure Start Time (Days into Year)	Failed Component	WIB Number	Duration (Hours)
308.465	Base Station 32	6	6.55
308.694	Base Station 15	5	1.50
308.698	Base Station 5	4	2.90
309.292	BSC-BC-Link 41	10	6.52

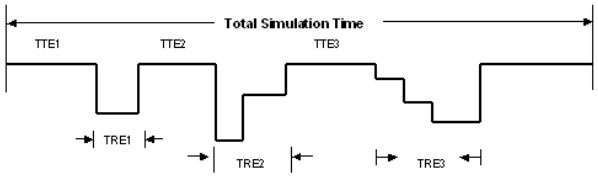


Fig. 5. Relationship of TTE, TRE and simulation time

Likewise, we expect the MTTE (mean of all times to epochs TTE), MTR (mean of all times to restore epochs TRE) and total simulation time to be:

$$MTTE = \frac{\sum_{i=1}^{n} TTE_i}{n} \qquad (6)$$

$$MTRE = \frac{\sum_{i=1}^{n} TRE_i}{n} \qquad (7)$$

$$Total\_Simulaiton\_Time = (MTTE + MTRE) \bullet n \qquad (8)$$

IV. RESULTS

As expected, the number of impact epochs increases as the network expands in all three scenarios since newly added WIBs in a wireless infrastructure will contribute more component outages. Figure 6 illustrates the relationship between the total numbers of impact epochs at different network size for each scenario over a one-year interval. Remember, this also includes single outage epochs. The nominal and degraded scenarios both use nominal MTTF, therefore the expected number of single component failures in these two scenarios should be at the same level when the network size is small, such as 1 or 2 WIBs, since the number of impact epochs is approximately the same. As the network size increases, the nominal scenario has more impact epochs as compared to the degraded maintenance scenario since longer repair times mean fewer components online at any instant that can fail. As it turns out, the degraded case has less epochs, but more multi-outage epochs. Remember – a one WIB network serves 100,000 customers while a ten WIB network serves 1,000,000.

 ISBN:978-1-61208-113-7

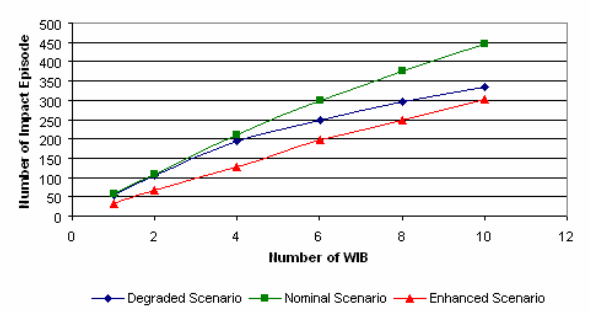


Fig. 6. Total number of impact epoch

Figure 7 displays the actual number of multi-outage epochs for each network size scenario. The curve increases almost linearly for networks in the degraded and nominal scenarios after network size exceeds 2 WIBs. The rate of growth slows down significantly in the enhanced scenario. Table 5 indicates that nearly 40% of the total impact epochs are multi-outage epochs in a 10 WIB network with the degraded scenario. This situation improves in the enhanced scenario, where less than 8% of total impact epochs include more than one outage.

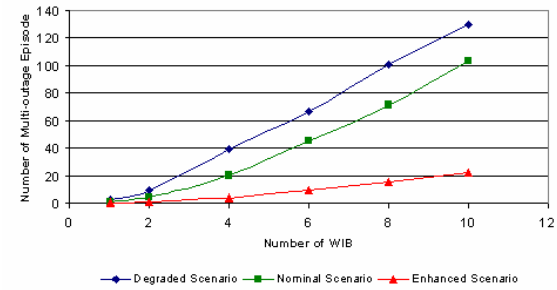


Fig. 7. Multi-outage epoch number

TABLE 5

Multi-Outage Impact Epoch Composition

# WIB	2 or more concurrent outages			3 or more concurrent outages		
	Degraded	Nominal	Enhanced	Degraded	Nominal	Enhanced
2	9.8%	4.6%	1.6%	1.1%	0.3%	0
4	20.1%	9.5%	3.3%	4.6%	1%	0
8	33.5%	18.3%	6.3%	12.7%	4.2%	0
10	39.5%	23.2%	7.7%	17.9%	6.5%	<0.9%

The difference between degraded and enhanced scenario is significant. The percentage of network epochs in the degraded scenario increases from 4.6% to 17.9% as it expands from 1 to 10 WIBs. The range is from 0.3% to 6.5% for networks in nominal scenario. While in an enhanced scenario network, the 3 or more outage epoch virtually disappears. Notable differences occur among three scenarios involving the multi-outage epochs. In the enhanced scenario, impact epochs consisting of more than 2 concurrent outages rarely happen, even when a network expands to serve 1 million customers. However, in the degraded scenario, when the network has 6 WIBs, the composition of impact epochs consisting of more than 2 concurrent outages is 7%. When the network has 10 WIBs, the number is 18%. Concurrent outages become a huge challenge for network operators in the degraded scenario, especially when network size grows.

The results of the network quiescent days for each scenario are shown in Figure 8. As the network expands, its quiescent availability decreases, almost linearly. In the degraded scenario, the total non-episodic time of a one WIB network is 345 days over a one-year operation time. By contrast, for a 10 WIB network, the number is only 213 days, which demonstrates that the network is in an episodic state 42% of the time. In the nominal scenario, which has the same reliability as the degraded scenario, the total non-episodic time of a 1-WIB network is 355 days, and 272 days for a 10-WIB network. This implies that 25% of the time the nominal network is in an episodic state for a 10 WIB network, which is approximately 30% improvement over the degraded scenario.

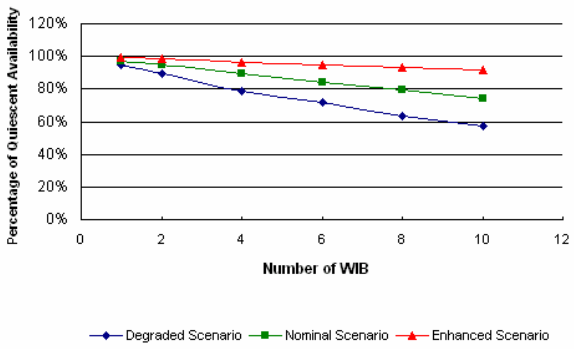


Fig. 8. Percentage of quiescent availability

The nominal and degraded scenarios use the same component reliability or MTTF. The difference is the component maintainability. Meanwhile, the nominal scenario is different from the enhanced scenario for both the component reliability and the maintainability. Figure 8, demonstrates that the nominal curve lies between the enhanced and degraded curves. Thus, the component maintainability rather than reliability is more decisive to the network quiescent availability. Efficient management of maintenance resources seems to have a positive impact on sustaining a network and avoiding an episodic status.

Figure 8 shows the quiescent availability of a network in different scenarios. There are four important attributes of an impact epoch: MTTE, MTRE, PCI, and WPLLH. MTTE is the average time between two impact epochs, which is used to model the network's reliability. MTRE is the average time to repair an impact outage in the network, which is a measure of the network's maintainability. PCI and PLLH are used to model the wireless network's survivability.

A. Mean Time To Epoch and Mean Time to Restore Epoch

Results demonstrate that MTTE decreases nonlinearly, as expected, as the network size increases for each scenario. In

all three scenarios, MTTE decreases quickly as the network grows from 1 to 3 WIBs, and the rate of decrease slows after 3 WIBs. The MTTE in degraded and nominal scenarios are very similar, as they have the same reliability. This is because single component outages are still dominant when the network is less than 3 WIBs. After that, as the network size increases, the overlapping phenomenon begins to play an important role in determining the total number of impact epochs.

MTRE is expected to increase as outage overlapping occurs. How much overlapping affects MTRE depends upon the pattern of the overlapping. There are several different overlapping patterns that could occur, shown in Figure 9 A, B, C, D. Among these four patterns shown in Figure 9, pattern "A" does not increase TRE since repair time of the second outage totally occurred within the repair time of the first outage (TRE in pattern "A" equals to the MTR of component one). Pattern "B" has a small degree of overlap and effect on TRE while pattern "C" has a moderate impact on TRE. Pattern "D" overlap is nearly sequential, having the largest impact on TRE. All these types of overlapping patterns may impact MTRE. Figure 10 illustrates the simulation output of the MTRE changes due to network size.

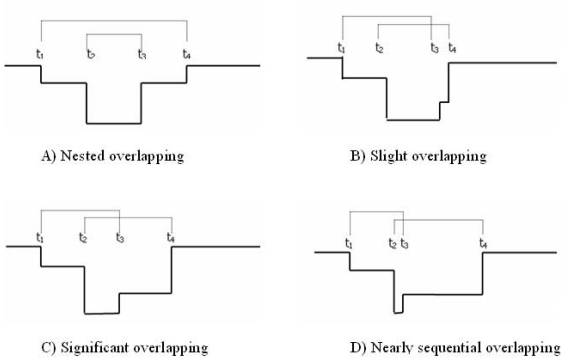


Fig. 9. Different Overlapping Patterns

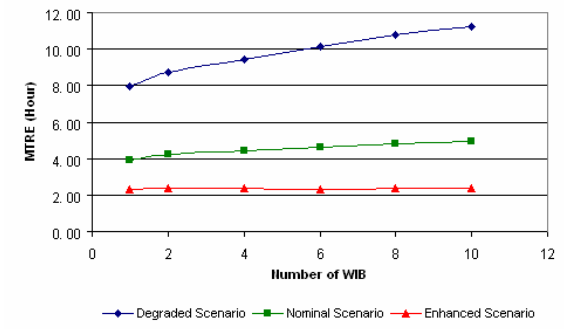


Fig. 10. MTRE in hours

As expected, MTRE in the degraded maintainability scenario increased nonlinearly as the network expanded due to overlapping outages. As the network grows, more overlapping instances occurred and the chance of overlapping pattern "A" increased, thereby decreasing MTRE. The component maintainability in the degraded scenario is lower than that in nominal and enhanced scenarios. The MTRE of a 10 WIB network in the degraded scenario increased by approximately 28% (about 144 minutes) from the single WIB network, while a 10 WIB network in the enhanced scenario increased by only 5.4 minutes longer than the one WIB network.

B. Peak Customers Impacted

A question that a network operator may ask is "what is the chance an impact epoch affecting more than 10,000 customers will occur in the next 30 days?" Understanding the distribution of peak customers impacted can provide insights into such questions. The PCI for each simulation run was collected and the data was fitted to an Exponential Distribution [2] with a high degree of significance (p value = 0.0000). This allowed easy calculation of probabilities of peak outages. Table 6 displays the probability of a PCI greater than or equal to 10,000 customers in 30 days for different scenarios and network sizes, along with the same results for a PCI greater than or equal to 5,000 customers. Larger networks have higher probabilities due to the additive nature of outages in epochs.

Table 6

Probability of PCI over 10,000 and over 5,000 customers in 30 days

Scenario Name	Number of WIB (over 10,000)				Number of WIB (over 5,000)			
	2	4	8	10	2	4	8	10
Degraded	3.3%	4.8%	10.3%	11.1%	18.2%	21.9%	32.1%	33.3%
Nominal	1.0%	1.1%	2.5%	2.7%	10.0%	10.6%	15.7%	16.4%
Enhanced	1.0%	1.1%	2.3%	2.4%	10.0%	10.7%	15.1%	15.6%

Similarly, the distribution of WPLLH values for networks of different sizes and scenarios are illustrated in Table 7. These results can predict the probability of PLLH over a threshold for a given time-period.

Table 7

WPLLH mean

Scenario Name	Number of WIB			
	2	4	8	10
Degraded	13,867	18,367	25,094	25,367
Nominal	6,409	6,640	8,088	8,257
Enhanced	3,550	3,735	4,042	4,506

The chance of the PCI and the PLLH over a certain threshold is much higher in the degraded scenario than that in the nominal and enhanced scenarios. For example, the chance of an epoch in which the PCI is over 10,000 customers over 30 days in the degraded network is three to five times than that of the enhanced scenarios. Thresholds are useful for network operators in effectively monitoring networks, given that they filter out lower priority epochs. In this paper, three different WPLLH threshold levels are used as filters: 5K WPLLH, 10K WPLLH and 15K WPLLH. A 5K WPLLH denotes that the product of impacted customers and impacted duration in an epoch is 5,000. For example, it

 ISBN:978-1-61208-113-7

could mean 5,000 customers are impacted for one hour or it could signify that 10,000 customers are impacted for half an hour. Figure 11 indicates the relationship between the numbers of impact epoch versus different thresholds, for the degraded scenario.

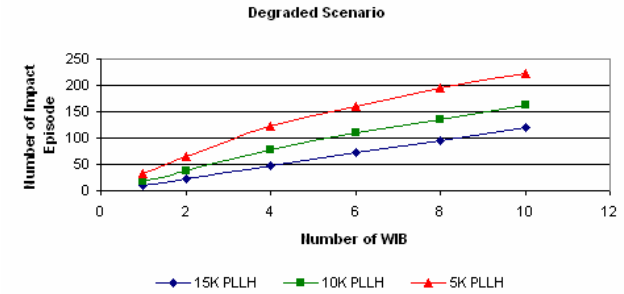


Fig. 11. Number of impact epochs with filters in degraded scenario

The growth rate of impact epochs over 5K WPLLH in all three scenarios increased rapidly as the network expands in size. At the size of 10 WIB, in the enhanced scenario, the number of impact epochs over 5K WPLLH is 52, while in the degraded scenario, the number is 223 (4 times more than enhanced scenario

This implies that in any scenario where a network expands, the number of impact epochs over a lower threshold can be expected to grow quickly. A network in the degraded scenario has to deal with a large number of epochs over higher thresholds because they grow in number at a much faster rate than that in the enhanced scenario. These insights should aid in network operators' ability to set efficient thresholds. Set too low, a threshold masks important outages; set too high, too many less significant outages are seen.

V. CONCLUSION

This work indicates that in large networks, the epoch perspective is useful in understanding the complex nature of ongoing concurrent failures. With these new metrics, operators can calculate such things as the probability of 3-outage epoch over a time-period and the probability of an epoch exceeding a specified peak over a time-period. Such information is useful to operators in allocating resources. Significant contributions of this work include:

- Defined the impact epoch as a new way to evaluate wireless network infrastructure's dependability.
- Developed new metrics for analyzing ARMS for large networks (MTTE, MTRE, Quiescent Availability, PCI and WPLLH).
- Development of empirically derived wireless traffic profiles to determine number of customers impacted by component failures by time of day and day of week.

Important conclusions include:

- An impact epoch perspective gives key insights into network dependability. Lacking empirical outage data, these perspectives are best investigated with simulation.
- Component maintainability has a large effect on a network's quiescent availability. Effective monitoring and efficient management of repair resources can shorten the time when a network is in an episodic state.
- The no. of small network impact epochs is not critical.

With respect to the last point, network operators should be very careful when expanding their infrastructure in order to accommodate more customers. Results here indicate that the number of concurrent outage epochs is sensitive to component reliability or maintainability. Reliability and maintainability should not be degraded in the expanded network. Additionally, it may be necessary to increase reliability and/or maintainability in order to keep multi-outage epochs to a minimum.

REFERENCES

[1] Albaghdadi M, and Razvi K (2005). Efficient transmission of periodic data that follows a consistent daily pattern. 9th IFIP/IEEE International Symposium on Integrated Network Management. IEEE Operations Center. Piscataway NY. pp 511-526.

[2] BestFit 4.5, Palisade Corporation.

[3] Cankaya H, Lardies A, and Ester G. (2004) Availability-aware analysis and evaluation of mesh and ring architectures for long-haul networks. Applied Telecommunications Symposium. pp. 116 – 121.

[4] Carver Carol Y, and Snow Andrew P (1999) Assessing the Impact of a Large-Scale Telecommunications Outage. Proceedings of the 7th International Conference on Telecommunication Systems. March: 483-489.

[5] Dryburgh L, and Hewett J (2004) Signaling System No.7: Protocol, Architecture and Services. ISBN 1-58705-040-4

[6] FCC Report (2005a) Report and Order and Future Notice of Proposed Rule Making, retrieve from http://www.fcc.org in August.

[7] FCC report, (2005b) http://www.fcc.gov/911/enhanced/, retrieved in August.

[8] Knight J, Strunk E, and Sullivan K. (2003) Towards a Rigorous Definition of Information System Availability. Proceedings of the DARPA Information Survivability Conference and Exposition. IEEE.

[9] Laprie J C (1995) Dependability – Its Attributes, Impairments and Means in Predictability Dependable Computing Systems.

[10] Leemis L (1995) Reliability Probabilistic Models and Statistical Methods. Prentice Hall, Englewood Cliffs, NJ.

[11] Lewis E E (1987) Introduction to Reliability Engineering. ISBN 0-471-81199-8.

[12] Malloy A D (2002) Design and Performance Evaluation of QoS-Oriented Wireless Networks. Georgia State University, Doctorial Dissertation.

[13] Snow A (1998) A Survivability Metric for Telecommunications: Insights and Shortcomings. 1998 Information Survivability Workshop – ISW'98 IEEE Computer Society, FL, October. pp.135-138.

[14] Snow Andrew P (2003) The Failure of a Regulatory Threshold and a Carrier Standard in Recognizing Significant Communication Loss. TPRC 2003. November.

[15] Snow A P (2004) Assessing pain below a regulatory outage reporting threshold. Telecommunications Policy, Vol. 28/7-8. pp 523-536.

[16] Snow A, Varshney U, and A Malloy (2000) Reliability and Survivability of Wireless and Mobile Networks. IEEE Computer Magazine. July. pp. 49-55.

[17] T1A1.2 Working Group (1997) A Technical Report on Network Survivability Performance

[18] WCB, (2005) The Wireless Industry and Its Contribution – A Presentation to Wire-line Competition Bureau.

 ISBN:978-1-61208-113-7

Soft Error Mask Analysis on Program Level

Lei Xiong, Qingping Tan, Jianjun Xu
School of Computer
National University of Defense Technology
Changsha 410073, China
leixiong@nudt.edu.cn, eric.tan.6508@gmail.com, jjun.xu@gmail.com

***Abstract*—Computer hardware which consist billions of transistors could fail because of soft errors with the improving of semiconductor technology, and these failure could result in incorrect program execution. However, soft errors could be masked. Most methods of SIFT (Software-Implemented Fault Tolerance) do not take mask into account. In fact, these parts of program which could mask soft errors don't need to be added redundancy instructions with SIFT methods. These parts of program are analyzed in this paper. We equal logical errors of soft errors to errors in program. In our analysis, we focus on two parts of program. One is idle program which is related to control flow. The probability for them to be executed is low. The other is dynamically dead codes. From static view, dynamically dead codes include statically dead codes and statically partial dead codes. By control flow graph, we analyze the condition which could mask soft errors of these parts of program. Finally, we design an experimental framework to demonstrate our analysis. Those codes which we analyze show their ability to mask soft errors.**

Keywords-soft errors; error mask; fault tolerance; control flow; dynamically dead codes

I. INTRODUCTION

Due to improvement of semiconductor technology, transistors are getting smaller and faster. Those transistors yield performance enhancements, but their lower threshold voltages and tighter noise margins make them less reliable. When facing energetic particles striking the chip, microprocessors which consist billions of transistors are susceptible to transient faults. Transient faults which are also called soft errors are intermittent faults. These faults do not cause permanent damage, but may result in incorrect program execution by altering transferring signals or stored values [1][2][3].

When soft errors appear, system could not be failure because of soft error mask. For an instance, when the hardware component which is not used encounters soft error, it can't affect system state. Soft error mask can be classified several classes based on its effects on different level, such as mask on architecture level and mask on program level. The mechanism of soft error mask on program level is that program errors which are caused by soft errors of hardware are masked on the level of program with control flow and data flow.

SIFT methods protect program to tolerate soft errors. Most of SIFT methods are implemented by compiler. The compiler which are implemented tolerance algorithm compile common program to redundancy program. These methods copy data which are used in program, and compute every part of program twice to make sure the right result. Most methods of SIFT do not take soft error mask into account. However, it is obviously that there is no need to copy those data which are related to those parts of program which can mask soft errors. If we can distinguish these parts of program which could mask soft errors from the whole program, we can ignore these parts of program with a fault tolerance method. Then, the method can lead to higher performance.

This paper gives an analysis of soft error mask on the level of program. Based on our analysis, we show these parts of program which could mask soft error. In accordance with those methods that protect program statically, these parts of program are analyzed from a static approach. Those parts of program which are idle program and dynamically dead codes can mask program errors which are caused by soft errors. Idle program which have low probability to be executed can't activate their program errors. Dynamically dead codes on the contrary are executed, but their results don't come to be used. To meet the trade-off between reliability and performance, we give a measurement to these parts of program to decide if these parts of program need to be protected with our method. To demonstrate our analysis, we give some experiments by the method of simulation. Our rest structures are as follows. In Section 2, we show two parts of program which could mask error. Section 3 analyzes conditional branches which is one of mask parts in program. Section 4 shows dynamically dead codes which is another mask part of program. We give an experiment frame in Section 5. We show results of experiments and discuss these results in Section 6. Section 7 is related work. In last Section, we make a conclusion.

II. SOFT ERRORS AND ITS MASK ON PROGRAM LEVEL

Radiation of cosmic rays has an effect on semiconductor chip. This event can cause a single event upset (SEU), and SEU cause soft errors of hardware. These hardware include storage, bus, cache and functional unit related to instruction pipeline. Soft errors of hardware could lead to errors of program during execution. These program errors could

 ISBN:978-1-61208-113-7

propagate in program with control flow and data flow. We consider program errors which are caused by soft errors as logical errors of soft error. Based on the mechanism that soft errors on hardware affect program execution, we choose to show program behaviors on the low level. It is convenient to show the effect of soft errors to program on low level, such as assembly language or machine language.

Based on our ground, when facing on energetic particle striking, inter-arrival times for raw faults in hardware components are independent and exponentially distributed with density function $\lambda e^{-\lambda t}$ [4]. When time t, in program, we assume that the instruction whose execution is related to time t is in the location of l. The number of functional units which are related to execution of the instruction is denoted as n. Density functions of these hardware components are denoted as $\lambda_1 e^{-\lambda_1 t}$, $\lambda_2 e^{-\lambda_2 t}$, $\cdots$, $\lambda_n e^{-\lambda_n t}$. The event that result of the instruction execution is right equals to the event that these n functional units are reliable. We denote the error probability of the instruction in the location of l as $P(l)$. Then

$$P(l) = 1-(1-P(u_1))(1-P(u_2))\cdots(1-P(u_n)) \quad (1)$$

and

$$P(l) = \int_0^t (\lambda_1 + \lambda_2 + \cdots + \lambda_n) e^{-(\lambda_1+\lambda_2+\cdots+\lambda_n)t} dt \quad (2)$$

We denote $\lambda_S = \lambda_1 + \lambda_2 + \cdots + \lambda_n$, so

$$P(l) = \int_0^t \lambda_S e^{-\lambda_S t} dt \quad (3)$$

We transform this formula:

$$P(l) = \int_0^t \lambda_S e^{-\lambda_S t} \frac{dt}{dl} dl \quad (4)$$

Since execution time of most instructions is single cycle, except some special instructions which need additional instruction cycles. To simplify our model, we assume that the average execution time of every instruction is the same. As a result, time and number of executed instructions keep linear relationship. So

$$P(l) = \int_0^l \lambda_S e^{-\lambda_S l} dl \quad (5)$$

We can see from this formula that error locations in program are almost exponentially distributed.

If soft errors of hardware lead to wrong results of instructions and wrong results don't affect program outcome, we take the situation as soft error mask. Soft error mask is a dynamic conception which is related to execution of program. However, dynamic behaviors of program are based on their static program. In this paper, program is modeled statically. We discuss program from static approach. There are several situations can mask soft errors which are as follows.

1) Idle program can mask errors. We define idle program as the part of program which are not executed during one execution. These program parts which can't be executed, they are pure idle program. The quantity of this program is small. Additionally, those parts of program which have probability to be executed can be idle program. These parts of program are not executed during one execution. Soft errors on hardware which are related to these parts of program are masked.
2) Dynamically dead instructions can mask errors in program. The results of dynamically dead instructions may not be used. They include two parts of program in static program. One is statically dead codes, their results are not used after these code. So their errors can be masked. The other one is partially dead codes. Partially dead codes are related to control flow. On some paths results of these codes are used, but on some other paths their results are dead. If results of these codes are dead definitely, their errors are masked.

III. ERROR MASK OF IDLE PROGRAM

Idle program are parts of program which can't be executed in one execution. We use control flow graph to describe the control flow of program. Its node represents basic block which contains sequential instructions, its edge represents a possibility of control flow path transition. Edge decides the block which follows the last block.

A. Idle Program

We assume the process that one node chooses the next node is a Markov process. Markov process is a type of stochastic process in which the outcome of a given trial depends only on the current state of the process. A system consisting of a series of Markov events is called a Markov chain [8]. As Figure 1 shows, we assign every basic block a label, such as A_1, A_2, $A_3, \cdots, A_n$, and we divide them into several groups based on their position. We denote them as G_i. For example, A_1 is the first node, $G_1 = \{A_1\}$, A_2, A_3, A_4 can be the next node of A_1, so we group them together, $G_2 = \{A_2, A_3, A_4\}$. Like this, we group A_5, A_6 together, A_7, A_8, A_9 are single group, $G_3 = \{A_5, A_6\}$, $G_4 = \{A_7\}$, $G_5 = \{A_8\}$, $G_6 = \{A_9\}$. We denote the basic block which is chosen to be executed as A_{ci} according to G_i, and $A_{ci} \in G_i$. Moreover, we denote the set of chosen basic block as $S_{ci} = \{A_{c1}, A_{c2}, A_{c3}, \cdots, A_{c(i-1)}\}$. According to the definition of Markov chain above, for a program, $S_{ci} (i = 1, 2, 3, \cdots, n)$ is a Markov chain. Every S_{ci} is only decided by the state of $S_{c(i-1)}$. It is independent of other states. We denote the probability of every branch as P. If the branch between basic block A_i and basic block A_j, we denote P_{ij} as probability of the branch. We define the probability of a branch as the ratio of the branch executing times to all executing times of source basic block. If the branch is definitely executed, the probability is 1, but we also describe it as P_{ij}.

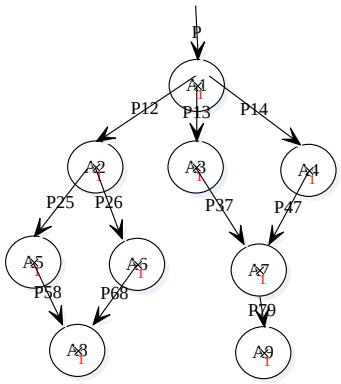


Figure 1. Control flow graph

B. *Probability Tree and Its Effects to Soft Error Mask*

As control flow graph shows, there is a probability for a node to next node. And next node also has probability to choose its next node. With control flow, every branch has its probability to be executed. We call it as probability tree. The probability of every branch is basic unit of this tree.

We analyze the probability tree from top to down. We assume that the real probability of the first node is P. We assume that the branch is from basic block i to basic block j. The real probability of this branch can compute like this: $P_{ij(real)} = P_{(basicblock)i} * P_{ij}$. The real probability of this branch from basic block i to basic block j is the multiplicative result of probability of basic block i and probability of this branch. The execution probability of basic block i is the sum of branch probability which points to the basic block i. This is: $P_{(basicblock)i} = \sum P_{*i(real)}$. We know the real probability of the first basic block, so we can compute all the real probability of probability tree from top to down.

To evaluate the trade-off between performance and reliability of a tolerance computer system, we must not over care the reliability of the computer system. So we have reasons to ignore protection to some parts of program. We give a threshold probability t. The threshold probability is suitable for system requirement that meets the trade-off between performance and reliability. We propose a new software-implemented fault tolerance method which pursues the trade-off between performance and reliability of system. In our method, branch probability which is less than t is ignored to be protected. In control flow graph, if the node which is pointed by a branch is an end node, then there is no other basic block next this basic block, we ignore the destination node to be protected. If the node is not an end node, there is a sub tree began with the destination node, we ignore this sub tree to protect. For a static program, after the application of our method to tolerate fault, the program which we will protect is transformed. To the transformed program, we do not need to adjust the probability of branches. For example, in Figure 1, if $P_{12} < t$, we ignore A_2 with its next branches and nodes to be protected. As a result, the program slice that we will protect is transformed. The program slice is described as Figure 2.

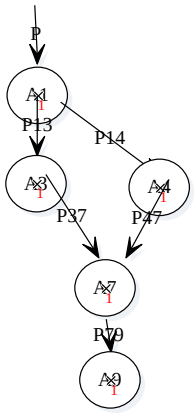


Figure 2. Control flow graph of program slice

IV. ERROR MASK OF DYNAMICALLY DEAD CODES

Dynamically dead instructions are those instructions whose results are dead when program runs. They include two classes of static codes. One is statically dead codes. The saved results of this kind of codes are not used before the saving place was written again. Shown as Figure 3, I2 is statically dead instruction. The other one is partially dead instructions. There are more than one path after this instruction. On some execution paths the saved results of this kind of instruction were not used before the saving place was written again. The instruction is dead on those execution paths. While, on other execution paths, the saved results of this kind of instruction are used. On these execution paths the instruction is alive. Partially dead instruction is also related to conditional branch. Figure 4 shows an instance of partially dead instruction. I2 assigns a value to variable X. After I2, there are two branches. Variable X is used on the left branch, but it is not used before assigning a new value to variable X of I5 on the right branch. So instruction I2 is alive if left path is chosen, and instruction I2 is dead if right path is chosen.

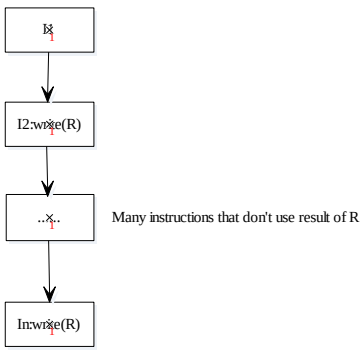


Figure 3. Statically dead instructions

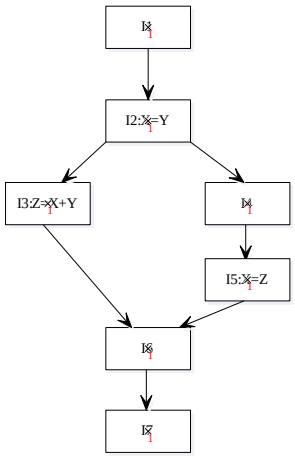


Figure 4. Partially dead instructions

If an instruction is partially dead instruction, it must be in a basic block which contains this instruction. Then the execution probability of this instruction equals to the execution probability of the basic block, and the latter has been computed in Section 3. Then $P_{PDI} = P_{basicblock}$. After this partially dead instruction, there are many paths, some make this instruction live, and some make this instruction dead. We can compute the sum probabilities of which make the partially dead instruction live. It equals to the sum of branch probability which can make the instruction live, and the sum probability of partially dead instruction to live is $\sum P_{live}$. So the final probability of this partially dead instruction to live is P_{1-live} , $P_{1-live} = P_{PDI} * (\sum P_{live})$. We assume a threshold probability as p . The variable p also meets the extent of trade-off between performance and reliability. To every partially dead instruction, if $P_{1-live} < p$, we ignore protection to this partially dead instruction. However, if $P_{1-live} > p$, we still have to protect it with redundancy instructions. As Figure 4 described, if $P_{1-live} < p$, the program slice, which we will protect, is transformed. The program slice is as Figure 5.

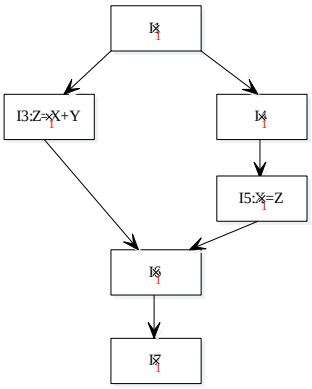


Figure 5. Control flow graph of program slice

.

V. EXPERIMENTAL FRAMEWORK

To demonstrate our analysis, we design an experimental framework. Our experimental framework includes SPEC2000 benchmarks, simplescalar tool set, and static fault injection tool. We choose some integer benchmarks from the SPEC2000 benchmark suite: gzip vpr, gcc, mcf, crafty, parser, perlbmk, gap, vortex, bzip2, twolf. Their sources are all written by C language. These benchmarks were compiled using a modified version of GNU GCC2.7.2.3 at the –O2 optimization level. To evaluate the mask ability of benchmark, we run benchmark on simplescalar tool set. Simplescalar tool set is a simulation tool for simulating computer architecture of a processor. We use sim-safe functional simulator to run our benchmark. Sim-safe function simulator is safer compared with other functional simulator such as sim-fast. We implement a static fault injection tool to inject fault into benchmark program. This tool can inject fault to determined location in program, and it can also inject fault to any random location in program.

The experimental framework is described as Figure 6. Firstly, we compile benchmarks to assembly codes. Then, we inject faults into assembly codes with fault injection tool. Thirdly, we run this program on simplescalar tool set with sim-safe simulator. Finally, we get simulation results.

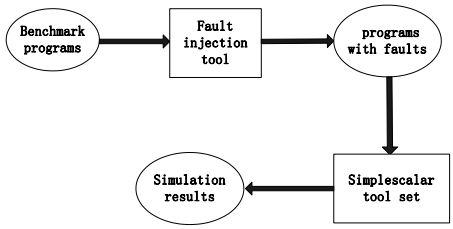


Figure 6. Experimental framework

VI. RESULT AND DISCUSSION

We give three error injection experiments. First, we inject errors into determined locations which contain idle program and partial dead codes. Second, we randomly inject error into locations which contain idle program and partially dead codes. Third, we randomly inject error into locations which are all over the whole program. The first experiment is to show that if the result of program can be still right when these code encounter faults. The second experiment is also to show this aim when faults are random. The third experiment show results under the real situation that faults turn on random location of program. In Section 2.1, we have known error locations in program are exponential distributed. In our experiment, we generate data by exponential distribution with Monte Carlo simulation.

 ISBN:978-1-61208-113-7

The statistics of conditional branch in SPEC2000 benchmark is as table 1. We can see from table 1 that the quantity of conditional branch instructions is much enough to be paid attention. Take benchmark bzip2 for example, the whole number lines of its code is 4650 lines, the number of its "if" statement is 245, and the number of its "switch" statement is 10. One "if" statement is at least one line, and one "switch" statement is at least two lines. These conditional branch instructions are at least 265 lines of code. Moreover, if we only care 80% of them, the rest codes which we need not to protect is at least 53 lines. However, these codes could be double or even more than 53 lines. If the scale of program is larger than bzip2, these codes could be more.

TABLE I. STATISTICS OF CONDITIONAL BRANCH IN SPEC2000

Spec name	Counts of "if" statement	Counts of "switch" statement
gzip	400	2
vpr	870	38
gcc	12805	524
mcf	80	1
crafty	1098	29
parser	624	0
perlbmk	4174	183
gap	2921	23
vortex	3018	49
Bzip2	245	10
twolf	1378	5

The percentage of statically partial dead code with SPEC2000 is as Figure 7. We can see from the figure that partial dead code is an important part of program. Among 11 SPEC2000 benchmarks, the percentage of statically partial dead code varies from 0.1 to nearly 7. In additional to idle program, these parts of program take a lot of account of the program. With our software-implemented fault tolerance method to tolerate soft errors, there could be a dramatic enhancement to system performance.

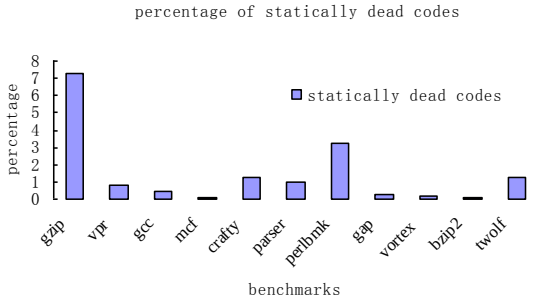


Figure 7. Percentage of statically partial dead codes in SPEC2000

We choose one of the SPEC2000 benchmark gzip to show our experimental results. For the ability of soft errors mask, gzip can present the behavior of all programs. Gzip is a data compression program. It uses Lempel-Ziv coding (LZ77) as its compression algorithm. We choose 1000 files to be compressed with gzip, and then we calculate the probabilities of conditional branches. We choose 20% as a threshold probability. It means we only care about those conditional branches whose probability is above 20%. Based on the frequency of soft errors and our program length, we statically inject 5 faults into the program in every experiment. For each experiment, we test 10 times. Then we calculate the average results. Results of Experiment are as Figure 8.

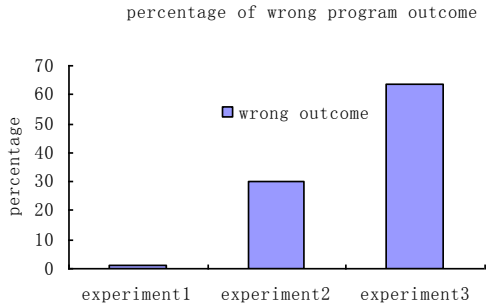


Figure 8. Simulation results of gzip with fault injection

Experiment 1 shows fault tolerance ability of program part which has low probability to be executed. Experiment 2 shows fault tolerance ability of program part which has probability to be executed. Experiment 3 shows fault tolerance ability of the whole program. From results of experiment 1, when there are errors in conditional branches or statically partial dead code whose probability to be executed is low, execution of the program can not be affected. This result matches our analysis. From Figure 8 of experiment 2, the effects of conditional branches and statically partial dead code to the program outcome is limited, these effects are weakened by control flow. From experiment 3, we can see that the ability of program to tolerate errors is considerable during its execution. Except these codes which are related to control flow can tolerate errors in program, there are still other codes can tolerate errors, such as errors are masked from logical operation.

VII. RELATED WORK

Soft error mask has been studied on the level of architecture. Mukherjee distinguishes these bits which do not affect program outcome as unACE bits [1]. If there is a fault on unACE bits, there is no effect to the system because unACE bits can't affect committed architectural state. Mukherjee also proposes AVF (Architectural Vulnerability Factor) to measure reliability of hardware component. AVF is the probability that a user-visible error will occur given a

 ISBN:978-1-61208-113-7

bit flips in a storage cell. LI et al. have shown AVF of hardware component is a variable because of soft error mask [4]. Mukherjee et al. show several situations of unACE bits on micro-architecture and architecture level [1].

There are some algorithms of soft error tolerance on software level, such as EDDI, SWIFT, and so on [2][3][5][6]. They are different from the method implemented by hardware duplication. Hardware methods protect hardware structure and duplicate hardware structure which can be corrupted. However, software methods protect programs by duplicating instructions and adding checking instructions into program [5] [6]. We take EDDI for example. It is implemented by compiler. The compiler compile source program to executable program containing redundancy code. EDDI divides program into many basic blocks. Based on basic block, it partition basic block by store instruction, this is called storeless basic block. Inside a storeless basic every instruction is duplicated, and before store instruction every result is checked [5]. However, EDDI, SWIFT protect program on the uniform way. No matter whether these sections of program can mask error or not, they protect them on the same way. Although these methods can get a good reliability of system, they sacrifice the performance of system.

Except these soft errors mask on architecture level, there are some situations for soft error to be masked on the level of program. Based on software protection, we give an analysis of soft errors mask on program level. Our results of analysis are helpful to optimize these tolerance systems which are implemented by the method of software-implemented hardware fault tolerance. The purpose of our analysis is to meet the trade-off between performance and reliability. If these parts of program which can mask soft error are clear for us, there is no need to protect these parts of program. Therefore, redundancy instructions are reduced. Performance of system can be improved.

VIII. CONCLUSION

Model transistors of processor get smaller and faster, but their lower threshold voltages and tighter noise margins make them less reliable. When computer system expose in the space, its components may encounter soft errors because of high energy particle striking. Once soft error of computer component happened, it may affect the execution of program. However, soft errors of computer components may not lead to system failure, because these soft errors may be masked. This paper analyze soft error mask on program level. Based on the mechanism that soft errors affect system reliability, we compute error distribution caused by soft error in program. In our analysis, there are two kinds parts of program related to control flow can mask soft error. They are idle program which is related to conditional branches and dynamically dead codes. Based on control flow graph, the probability of branches and basic block to be executed are computed by our method. In our method, if the probability of basic block to be executed is less than threshold probability, we considered this basic block as idle program. These parts of program need not to be protected. Dynamically dead code includes statically dead code and partially dead code from static approach. In our method, partially dead code whose probability to live is less than threshold are ignored to be protected. We designed an experiment frame to demonstrate our analysis. In our experiment, we statically inject faults to program, and run program on a simulator. Experiment Results match our analysis.

REFERENCES

[1] S. Mukherjee, Architecture Design for Soft Errors, USA, Morgan Kaufmann Publishers, 2008.

[2] G.A. Reis, J. Chang, and N. Vachharajani, "Software-Controlled Fault Tolerance. ACM Transactions on Architecture and Code Optimization", Vol.V, No. N, 2005, pp. 1-28.

[3] G.A. Reis, "Software Modulated Fault Tolerance", A dissertation presented to the faculty of Princeton University, 2008.

[4] X. Li, "Soft Error Modeling and Analysis for Microprocessors", A dissertation presented to computer science in the graduate college of the University of Illinois, 2008.

[5] N. Oh et al., "Error Detection by Duplicated Instructions in Super-Scalar Processors", IEEE Transactions on Reliability, Vol. 51, No. 1, March 2002, pp. 63–75.

[6] G.A. Reis, J. Chang, N. Vachharajani, R. Rangan, and D. I. August, "SWIFT: Software-implemented Fault Tolerance", In Proceedings of the 3rd International Symposium on Code Generation and Optimization. March 2005, pp. 243–254.

[7] A. Benso, S.D. Carlo, and G.D. Natale, "Static Analysis of SEU Effects on Software Applications", International test conference. IEEE, 2002, pp. 500-508.

[8] S. Sparks, S. Embleton, and R. Cunningham, "Automated Vulnerability Analysis Leveraging Control Flow for Evolutionary Input Crafting", Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007), 2007, pp.477-486.

[9] J. L. Hennessy and D. A. Patterson, Computer Architecture: A Quantitative Apprach, Third Edition, BeiJing, China Machine Press, 2002.

[10] L. David and I. Pusut, "Static Determination of Probabilistic Execution Times", Proceedings of the 12th 16th Euromicro Conference on Real-Time System, ECRTS, 2004.

[11] J. Singer, "Towards Probabilistic Program Slicing", Dagstuhl Seminar Proceedings 05451 Beyond Program Slicing, 2006.

[12] M. Weiser, "Program slicing", IEEE Transactions on software Engineering. August 1984, pp. 352-357.

[13] J.A .Butts and G. Sohi, "Dynamic Dead-Instruction Detection and Elimination", In 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS). October 2002, pp. 199-210.

[14] B. Fahs, S. Bose and M. Crum, "Performance Characterization of a Hardware Mechanism for Dynamic Optimization", In 34th Annual International Symposium on Microarchitecture (MICRO). December 2001, pp. 16-27.

[15] A. V.Aho, R. Sethi, and J.D. Ullman, Compilers: Principles, Techniques and Tools, Addison-Wwsley, 1985.

[16] S.S. Muchnick, Advanced Compoiler Design Implementation, Elsevier, 1997.

[17] J. Xue, Q. Cai, and L. Gao, "Partial Dead Code Elimination on Predicated Region", software-practice and experience. 36, 2004, pp. 1655-1685.

 ISBN:978-1-61208-113-7