International Journal on Advances in Networks and Services

2012 vol. 5 nr. 3&4

The *International Journal on Advances in Networks and Services* is published by IARIA.
ISSN: 1942-2644
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Networks and Services
Volume 5, Number 3 & 4, 2012

Mahmoud Brahimi, University of Msila, Algeria
Marco Bruti, Telecom Italia Sparkle S.p.A., Italy
Dumitru Burdescu, University of Craiova, Romania
Diletta Romana Cacciagrano, University of Camerino, Italy
Maria-Dolores Cano, Universidad Politécnica de Cartagena, Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Eduardo Cerqueira, Federal University of Para, Brazil
Patryk Chamuczyński, TechniSat, Poland
Bruno Chatras, Orange Labs, France
Marc Cheboldaeff, T-Systems International GmbH, Germany
Kong Cheng, Telcordia Research, USA
Dickson Chiu, Dickson Computer Systems, Hong Kong
Andrzej Chydzinski, Silesian University of Technology, Poland
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Noelia Correia, University of the Algarve, Portugal
Noël Crespi, Institut Telecom, Telecom SudParis, France
Paulo da Fonseca Pinto, Universidade Nova de Lisboa, Portugal
Philip Davies, Bournemouth and Poole College / Bournemouth University, UK
Carlton Davis, École Polytechnique de Montréal, Canada
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
João Henrique de Souza Pereira, University of São Paulo, Brazil
Javier Del Ser, Tecnalia Research & Innovation, Spain
Behnam Dezfouli, Universiti Teknologi Malaysia (UTM), Malaysia
Mari Carmen Domingo, Barcelona Tech University, Spain
Daniela Dragomirescu, LAAS-CNRS, University of Toulouse, France
Jean-Michel Dricot, Université Libre de Bruxelles, Belgium
Wan Du, Nanyang Technological University (NTU), Singapore
Matthias Ehmann, Universität Bayreuth, Germany
Wael M El-Medany, University Of Bahrain, Bahrain
Imad H. Elhajj, American University of Beirut, Lebanon
Gledson Elias, Federal University of Paraíba, Brazil
Joshua Ellul, Imperial College, London
Rainer Falk, Siemens AG - Corporate Technology, Germany
Károly Farkas, Budapest University of Technology and Economics, Hungary
Huei-Wen Ferng, National Taiwan University of Science and Technology - Taipei, Taiwan
Gianluigi Ferrari, University of Parma, Italy
Mário F. S. Ferreira, University of Aveiro, Portugal
Bruno Filipe Marques, Polytechnic Institute of Viseu, Portugal
Ulrich Flegel, HFT Stuttgart, Germany
Juan J. Flores, Universidad Michoacana, Mexico
Ingo Friese, Deutsche Telekom AG - Berlin, Germany
Sebastian Fudickar, University of Potsdam, Germany
Stefania Galizia, Innova S.p.A., Italy
Ivan Ganchev, University of Limerick, Ireland
Miguel Garcia, Universitat Politecnica de Valencia, Spain
Emiliano Garcia-Palacios, Queens University Belfast, UK

Gordana Gardasevic, University of Banja Luka, Bosnia and Herzegovina
Marc Gilg, University of Haute-Alsace, France
Debasis Giri, Haldia Institute of Technology, India
Markus Goldstein, DFKI (German Research Center for Artificial Intelligence GmbH), Germany
Luis Gomes, Universidade Nova Lisboa, Portugal
Anahita Gouya, Solution Architect, France
Mohamed Graiet, Institut Supérieur d'Informatique et de Mathématique de Monastir, Tunisie
Christos Grecos, University of West of Scotland, UK
Vic Grout, Glyndwr University, UK
Yi Gu, University of Tennessee, Martin, USA
Angela Guercio, Kent State University, USA
Xiang Gui, Massey University, New Zealand
Mina S. Guirguis, Texas State University - San Marcos, USA
Tibor Gyires, School of Information Technology, Illinois State University, USA
Keijo Haataja, University of Eastern Finland, Finland
Gerhard Hancke, Royal Holloway / University of London, UK
R. Hariprakash, Arulmigu Meenakshi Amman College of Engineering, Chennai, India
Go Hasegawa, Osaka University, Japan
Hermann Hellwagner, Klagenfurt University, Austria
Eva Hladká, CESNET & Masaryk University, Czech Republic
Hans-Joachim Hof, Munich University of Applied Sciences, Germany
Razib Iqbal, Amdocs, Canada
Abhaya Induruwa, Canterbury Christ Church University, UK
Muhammad Ismail, University of Waterloo, Canada
Vasanth Iyer, Florida International University, Miami, USA
Peter Janacik, Heinz Nixdorf Institute, University of Paderborn, Germany
Robert Janowski, Warsaw School of Computer Science, Poland
Imad Jawhar, United Arab Emirates University, UAE
Aravind Kailas, University of North Carolina at Charlotte, USA
Mohamed Abd rabou Ahmed Kalil, Ilmenau University of Technology, Germany
Kyoung-Don Kang, State University of New York at Binghamton, USA
Omid Kashefi, Iran University of Science and Technology, Iran
Sarfraz Khokhar, Cisco Systems Inc., USA
Vitaly Klyuev, University of Aizu, Japan
Jarkko Kneckt, Nokia Research Center, Finland
Dan Komosny, Brno University of Technology, Czech Republic
Ilker Korkmaz, Izmir University of Economics, Turkey
Tomas Koutny, University of West Bohemia, Czech Republic
Evangelos Kranakis, Carleton University - Ottawa, Canada
Lars Krueger, T-Systems International GmbH, Germany
Kae Hsiang Kwong, MIMOS Berhad, Malaysia
KP Lam, University of Keele, UK
Birger Lantow, University of Rostock, Germany
Hadi Larijani, Glasgow Caledonian Univ., UK
Annett Laube-Rosenpflanzer, Bern University of Applied Sciences, Switzerland
Angelos Lazaris, University of Southern California (USC), USA

Gyu Myoung Lee, Institut Telecom, Telecom SudParis, France
Ying Li, Peking University, China
Shiguo Lian, Orange Labs Beijing, China
Chiu-Kuo Liang, Chung Hua University, Hsinchu, Taiwan
Wei-Ming Lin, University of Texas at San Antonio, USA
David Lizcano, Universidad a Distancia de Madrid, Spain
Chengnian Long, Shanghai Jiao Tong University, China
Jonathan Loo, Middlesex University, UK
Edmo Lopes Filho, Algar Telecom, Brazil
Pascal Lorenz, University of Haute Alsace, France
Albert A. Lysko, Council for Scientific and Industrial Research (CSIR), South Africa
Pavel Mach, Czech Technical University in Prague, Czech Republic
Elsa María Macías López, University of Las Palmas de Gran Canaria, Spain
Damien Magoni, University of Bordeaux, France
Ahmed Mahdy, Texas A&M University-Corpus Christi, USA
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Gianfranco Manes, University of Florence, Italy
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Moshe Timothy Masonta, Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa
Hamid Menouar, QU Wireless Innovations Center - Doha, Qatar
Guowang Miao, KTH, The Royal Institute of Technology, Sweden
Mohssen Mohammed, University of Cape Town, South Africa
Miklos Molnar, University Montpellier 2, France
Lorenzo Mossucca, Istituto Superiore Mario Boella, Italy
Jogesh K. Muppala, The Hong Kong University of Science and Technology, Hong Kong
Katsuhiro Naito, Mie University, Japan
Deok Hee Nam, Wilberforce University, USA
Sarmistha Neogy, Jadavpur University- Kolkata, India
Rui Neto Marinheiro, Instituto Universitário de Lisboa (ISCTE-IUL), Instituto de Telecomunicações, Portugal
David Newell, Bournemouth University - Bournemouth, UK
Armando Nolasco Pinto, Universidade de Aveiro / Instituto de Telecomunicações, Portugal
Jason R.C. Nurse, University of Oxford, UK
Kazuya Odagiri, Yamaguchi University, Japan
Máirtín O'Droma, University of Limerick, Ireland
Rainer Oechsle, University of Applied Science, Trier, Germany
Henning Olesen, Aalborg University Copenhagen, Denmark
Jose Oscar Fajardo, University of the Basque Country, Spain
Constantin Paleologu, University Politehnica of Bucharest, Romania
Eleni Patouni, National & Kapodistrian University of Athens, Greece
Harry Perros, NC State University, USA
Miodrag Potkonjak, University of California - Los Angeles, USA
Yusnita Rahayu, Universiti Malaysia Pahang (UMP), Malaysia
Yenumula B. Reddy, Grambling State University, USA
Oliviero Riganelli, University of Milano Bicocca, Italy
Patrice Rondao Alface, Alcatel-Lucent Bell Labs, Belgium
Teng Rui, National Institute of Information and Communication Technology, Japan

Antonio Ruiz Martinez, University of Murcia, Spain
George S. Oreku, TIRDO / North West University, Tanzania/ South Africa
Sattar B. Sadkhan, Chairman of IEEE IRAQ Section, Iraq
Husnain Saeed, National University of Sciences & Technology (NUST), Pakistan
Addisson Salazar, Universidad Politecnica de Valencia, Spain
Sébastien Salva, University of Auvergne, France
Ioakeim Samaras, Aristotle University of Thessaloniki, Greece
Luz A. Sánchez-Gálvez, Benemérita Universidad Autónoma de Puebla, México
Teerapat Sanguankotchakorn, Asian Institute of Technology, Thailand
José Santa, University of Murcia, Spain
Rajarshi Sanyal, Belgacom International Carrier Services, Belgium
Mohamad Sayed Hassan, Orange Labs, France
Thomas C. Schmidt, HAW Hamburg, Germany
Hans Scholten, Pervasive Systems / University of Twente, The Netherlands
Véronique Sebastien, University of Reunion Island, France
Jean-Pierre Seifert, Technische Universität Berlin & Telekom Innovation Laboratories, Germany
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Dimitrios Serpanos, Univ. of Patras and ISI/RC ATHENA, Greece
Xu Shao, Institute for Infocomm Research, Singapore
Roman Y. Shtykh, Rakuten, Inc., Japan
Salman Ijaz Institute of Systems and Robotics, University of Algarve, Portugal
Adão Silva, University of Aveiro / Institute of Telecommunications, Portugal
Florian Skopik, AIT Austrian Institute of Technology, Austria
Karel Slavicek, Masaryk University, Czech Republic
Vahid Solouk, Urmia University of Technology, Iran
Peter Soreanu, ORT Braude College, Israel
Pedro Sousa, University of Minho, Portugal
Vladimir Stantchev, SRH University Berlin, Germany
Radu Stoleru, Texas A&M University - College Station, USA
Lars Strand, Nofas, Norway
Stefan Strauβ, Austrian Academy of Sciences, Austria
Álvaro Suárez Sarmiento, University of Las Palmas de Gran Canaria, Spain
Masashi Sugano, School of Knowledge and Information Systems, Osaka Prefecture University, Japan
Young-Joo Suh, POSTECH (Pohang University of Science and Technology), Korea
Junzhao Sun, University of Oulu, Finland
David R. Surma, Indiana University South Bend, USA
Yongning Tang, School of Information Technology, Illinois State University, USA
Yoshiaki Taniguchi, Osaka University, Japan
Anel Tanovic, BH Telecom d.d. Sarajevo, Bosnia and Herzegovina
Olivier Terzo, Istituto Superiore Mario Boella - Torino, Italy
Tzu-Chieh Tsai, National Chengchi University, Taiwan
Samyr Vale, Federal University of Maranhão - UFMA, Brazil
Dario Vieira, EFREI, France
Natalija Vlajic, York University - Toronto, Canada
Lukas Vojtech, Czech Technical University in Prague, Czech Republic
Michael von Riegen, University of Hamburg, Germany

Joris Walraevens, Ghent University, Belgium
You-Chiun Wang, National Sun Yat-Sen University, Taiwan
Gary R. Weckman, Ohio University, USA
Chih-Yu Wen, National Chung Hsing University, Taichung, Taiwan
Michelle Wetterwald, EURECOM - Sophia Antipolis, France
Feng Xia, Dalian University of Technology, China
Kaiping Xue, USTC - Hefei, China
Mark Yampolskiy, Vanderbilt University, USA
Dongfang Yang, National Research Council, Canada
Qimin Yang, Harvey Mudd College, USA
Beytullah Yildiz, TOBB Economics and Technology University, Turkey
Anastasiya Yurchyshyna, University of Geneva, Switzerland
Sergey Y. Yurish, IFSA, Spain
Faramak Zandi, La Salle University, USA
Jelena Zdravkovic, Stockholm University, Sweden
Yuanyuan Zeng, Wuhan University, China
Weiliang Zhao, Macquarie University, Australia
Wenbing Zhao, Cleveland State University, USA
Zibin Zheng, The Chinese University of Hong Kong, China
Yongxin Zhu, Shanghai Jiao Tong University, China
Zuqing Zhu, University of Science and Technology of China, China
Martin Zimmermann, University of Applied Sciences Offenburg, Germany

International Journal on Advances in Networks and Services
Volume 5, Numbers 3 & 4, 2012

CONTENTS

The Impact of Multi-Outage Episodes on Large-Scale Wireless Voice Networks

Andrew P. Snow and Yachuan Chen
Ohio University
School of Information and Telecommunication Systems
Athens, Ohio
e-mail: asnow@ohio.edu; yc137604@ohio.edu

Gary R. Weckman
Ohio University
Department of Industrial and Systems Engineering
Athens, Ohio
e-mail: weckmang@ohio.edu

Abstract— **Large wireless network infrastructures experience concurrent or overlapping service outages due to equipment and link failures. The frequency, duration, and impact of such episodes are of interest to users and network operators alike. Here, a research project which investigates through simulation the characteristics of concurrent network outages in large wireless network infrastructures is presented. The dependability attributes used to gain a perspective on this issue are network reliability, availability, maintainability and survivability. To assess these attributes in this setting, a new term, called an "impact epoch", is introduced. Epochs are defined as single, concurrent, or overlapping outages in time, consisting of *n* different outages. A wireless network is expanded in size and epochs observed as the network grows. The new proposed metrics offer valuable insights into the management of restoration resources. Simulations proved invaluable in identifying multi-outage epochs, as well as modeling their occurrence, frequency, duration, and size – results which are analytically intractable for assessing large networks.**

Keywords – RAMS; network outages; simulation; survivability; reliability; maintainability; wireless network infrastructure

I. INTRODUCTION

The larger the network, the greater the challenge for operators. Networks are 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 customers—as well as accommodate new ones—could be adversely affected. Understanding the characteristics of concurrent outages as a function of network size, 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 customers 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. This phenomenon was first reported in [1], and this paper expands on and extends some of those preliminary findings. There is much interest in understanding the impact of outages. Hariri, et al [2] examined the impact of concurrent faults and attacks in large-scale networks, in particular the internet. However, the emphasis was on the effect of multiple transmission and switching outages to traffic, not predictions of the frequency of such phenomena. Alternately, Bassiri and Heydari [3] considered network survivability in the presence of regional outage scenarios. However, they concentrated on the effects of multiple switch and link outages in regional areas due to such phenomena as natural disasters, and also concentrates on traffic in internet environments. Recently, others invented an outage management portal to coordinate response to single outages [4]. However, no studies could be found that examined the probabilistic frequency and severity of concurrent outages. Prior published research has not considered how often multiple outage epochs occur in large-scale networks, how many simultaneous outage epochs can be expected, and how many users can be expected to be impacted for how long.

A. Dependability

Users count on networks. If a network is unreliable, hard to maintain, and has poor availability, it can hardly be deemed successful. Dependability has a number of different attributes. According to Avižienis, et al [5], 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 [6].

The higher the survivability, the better the 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 RAMS attributes (reliability, availability, maintainability, and survivability) of dependability.

B. Reliability

Reliability is a function of how often we might expect failure. Conversely, the formal definition of network reliability is the probability that it will perform its required functions over a specific period of time [7]. The reliability for a network, a network service, 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 [5]:

$$R(t) = e^{-\lambda t} = e^{-1/MTTF} \quad (1)$$

Where λ is expected failure rate and MTTF (mean time to failure) is the 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

Maintainability is a function of how fast we can expect to recover from a failure. Network maintainability is defined as the ability of a network to recover from failures [8]. Maintainability can be determined from the Mean Time to Restore (MTTR). Restore time is a random variable and typically consists of three parts – detection time, travel time to the outage location and the actual repair or replacement time. In this research, the lognormal distribution is used, as travel time plays an important role.

D. Availability

There are two forms of availability – instantaneous and average. Network instantaneous availability is defined as the probability that a network is ready for use when needed [8]. Average availability can be expressed as:

$$A = \frac{MTTF}{MTTF + MTTR} \quad (2)$$

Availability, being the fraction of time a network or network service is up, is a good metric to assess the state when the network is experiencing no problems due to failures.

E. Survivability

Availability is not always a good indicator of network dependability, as networks are very rarely "all-up" or "all-down". Rather, networks are "mostly-up"—or said another way, "fractionally-up". Survivability is a measure that can capture this phenomenon. Network survivability is defined as the ability of a network to provide services to most customers under partial failures. Snow [9] defined Prime Lost Line Hours (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).

$$NS = 1 - \frac{PLLH}{TLH} \quad (3)$$

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 [10], [11], [12].

The organization of this paper follows. In Section II the concept of impact epochs is introduced, which represent multiple outages in time. In Section III, wireless voice infrastructure is introduced. Additionally, equipment and link reliability and maintenance are quantified. Then architectural scenarios investigated in this paper are presented. In Section IV the paper research questions are presented and discussed, while Section V introduces the simulation model used to address the research questions, and the assumptions and limitations of the model. Lastly, Sections V and VII present the results and conclusions, respectively.

II. IMPACT EPOCH

This research examines episodes where multiple outages overlap in time. Single outages impact some fraction of users. When they are coincident, the impact increases and challenges network operators. 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 a state of no customers impacted to a state of having customers impacted. 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 single

impact epochs, consisting of two non-overlapping outages, is shown in Fig. 1 in the form of an epoch profile. Note that time is represented by the X-axis, and the Y-axis represents the percentage of customers served in the network. Each outage has a duration and a maximum impact.

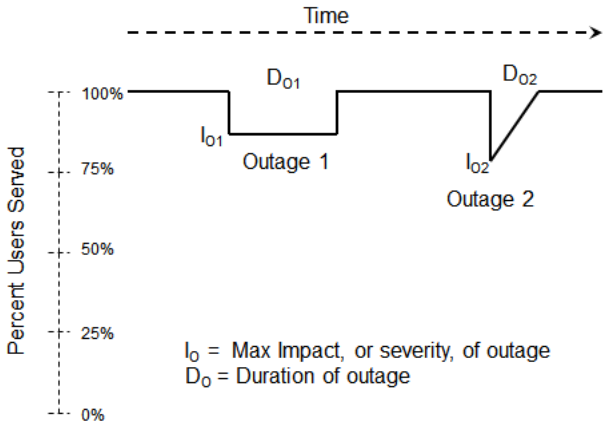


Figure 1. Non-Overlapping Outage Epochs

Next, refer to Fig. 2, which shows three different combinations of these two outages. Note how different these two events are, depending on degree of overlap. In the top profile, the two outages do not overlap and are separate epochs. In the middle profile, the outages combine into a single epoch with the same duration, but with a larger impact. Lastly, note the bottom profile, which has a different duration and impact.

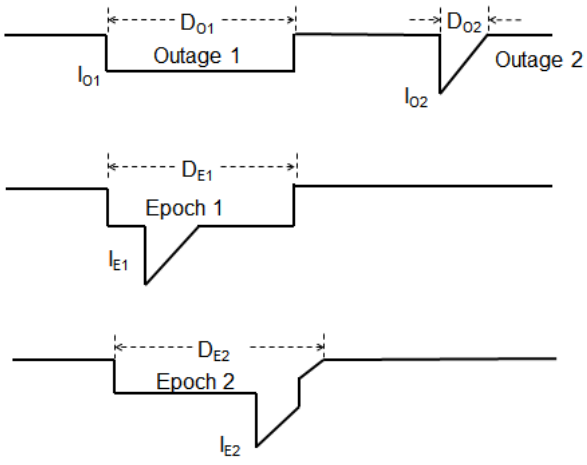


Figure 2. Different Perspectives of Two Overlapping Outages

Individual epochs are arrival events, and 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 lessen 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 recover gracefully 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:

$$A_Q = \frac{MTTE}{MTTE + MTRE} \quad (4)$$

Equation 3 can still measure survivability from an epoch perspective. However, in an environment where there may be concurrent or overlapping outages, peak customers impacted (PCI) may be of interest. For instance, in Fig. 2, epoch 2 has a larger PCI than epoch 1.

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).

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) Wireless Prime Lost Line Hours (WPLLH)

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 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.

III. WIRELESS NETWORKS

Like all telecommunications providers, wireless operators are reluctant to share statistics on service outages. Even so, 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 [13]. On the other hand, research in the world of wireless communication, especially in cell phone networks, is relatively new. Research into wireless telephone network reliability did not receive much attention until the late 1990s. Over the last 22 years, the wireless network has grown at an amazing rate. According to the Cellular Telecommunications Industry Association (CTIA) wireless Quick Fact Sheet [14], cellular subscribers in the US surpassed 5 million in 1990 and doubled in just two years. By 2012, cellular subscribers exceeded 300 million in the US and wireless penetration rate was over 65%. There were over 327 million customers in the US as of June, 2012.

In 1992, the FCC at first ruled that wire-line carriers had to report all outages that affected more than 50,000 customers for at least 30 minutes. This threshold was quickly lowered to 30,000 customers for 30 minutes in 1993 [10]. Thresholds for RAMS attributes have also been shown to be important in wireless networks [15]. 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 [16]. 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 [17].

A. Wireless Network Infrastructure

Wireless networks consist of components, such as cable and equipment. Additionally, equipment consists of both hardware and software. The general structure of a wireless network with most of the required functional components is shown in Fig. 3. 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) [18].

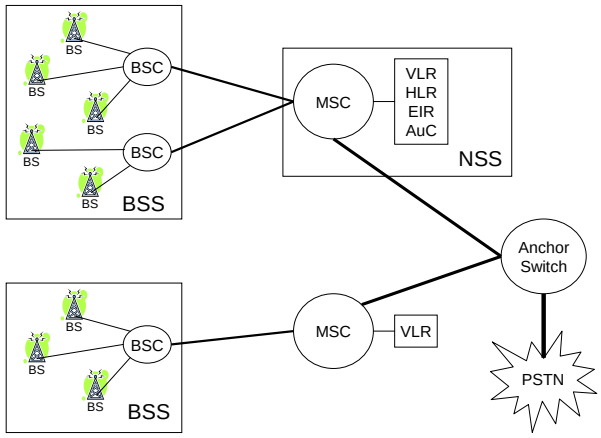


Figure 3. Wireless Network Infrastructure

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) [18]. 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, [19], 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 MTTR in a WIB might affect the network's dependability [19]. The topology used in a WIB is the star topology. Large wireless infrastructures consist of multiple WIBs.

B. Wireless Traffic

Wireless traffic, like all telecommunications traffic, varies widely over a single day. If equipment fails, or transmission links are severed, users are impacted. For faster restoration, providers use redundancy in equipment and links, and a topology that minimizes restoral times. Advantages of using the star topology include supporting modular expansion, as well as simplified monitoring and trouble-shooting. The largest disadvantage of star topology is the creation of a 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 [20].

TABLE 2. No. of Components in One WIB and Maximum Failure Impact

Component	Number in One WIB	Number 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	2,000
Anchor-MSC Link	1	100,000
Anchor Switch	*n*	*n* x 100,000
Anchor Link	*n*	*n* x 100,000

Note: *n* is the number of WIBs in the wireless infrastructure

The time factor accounts for time-of-day and day-of-week usage by customers. The goal of network engineering for carriers is to establish an infrastructure that satisfies peak hour traffic loading. Similar to traffic on a highway, voice traffic volume in networks varies over a day. According to historical statistics for wireline voice traffic [20], heavy traffic load in the wire-line network occurs between 9:00am and 4:00pm on weekdays. Taking traffic estimates into account, a network component failure occurring at different times may impact a different number of users. For example, a one-hour outage at 10:00am has much a larger impact than a one-hour failure at 3:00am in the morning. The time factor values, or utilization, for wire-line networks are summarized in Table 3 from [20].

TABLE 3. Time Factor for Wire-Line Network

Spanned	Time Period	Time Factor
Day	(8:00am to 4:59pm, Mon. ~ Fri.)	1.0
Evening	(5:00pm to 10:59pm, Mon. ~ Fri.)	0.3
Night	(11:00pm to 7:59am, Mon ~ Sun.)	0.1
Weekend	(8:00am to 10:59pm, Sat. & Sun.)	0.2

Say there is a failure of central office with 50,000 lines that lasts one hour. The number of affected customers is 1 x 50,000 = 50,000 if the outage started at 10:00am. However, if the outage started at 3:00am the number of affected customers is 0.1 x 50,000 = 5,000. The product of time factor and telecommunications capacity is the impact of the outage, in line hours. As the time factor are fractions of full utilization during the prime times of the day, this impact has been called prime lost line hours, or PLLH [9], [10], [11].

In this work, a new traffic profile for wireless networks is developed. This is because 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 the time of day a call is placed. For example, many cell phone plans offer free calls on weekends and after 9:00pm on weekdays. Some people could wait until 9:00pm to place calls and take advantage of this plan. Such phenomena results in different weekday and weekend traffic profiles in wireless networks. In Albaghdadi and Razvi [21], the authors studied an actual 1320 cell GSM network. In that research, the results reported in this GSM network were used to develop five-day weekday traffic and weekend traffic profiles as shown in Fig. 4 and 5 respectively. The data is from [21] while the solid lines are added for this research to create a wireless time factor. These wireless time factors were developed to create a wireless PLLH outage impact metric, called hereafter the WPLLH, where the W denotes wireless.

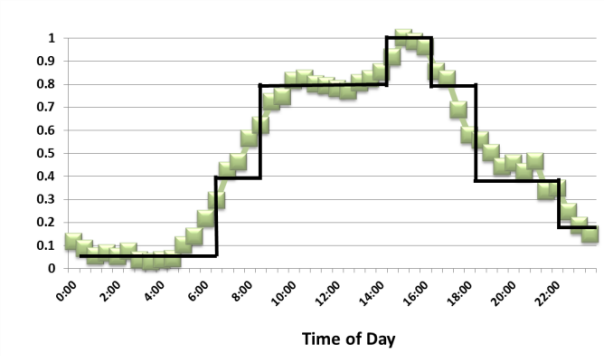


Figure 4. Wireless Weekday Time Factor

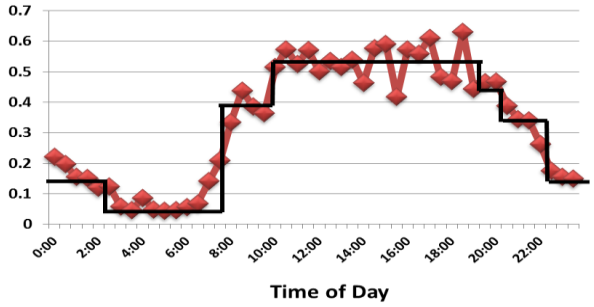


Figure 5. Wireless Weekend Day Time Factor

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 MTTRs).

C. Network Component MTTF and MTTR

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 MTTRs Used in the Study

Component Name	Nominal MTTF (Years)	Enhanced MTTF (Years)	Degraded MTTR (Hours)	Nominal MTTR (Hours)	Enhanced MTTR (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/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 [19]. 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.

Derivation of link MTTFs are also based upon empirical failure data for fiber optic links, and are derived here. As suspected, the MTTFs are greatly affected by power failures.

As seen in the multi-WIB architecture of Fig. 3, transmission systems include BS-BSC links, BSC-MSC links, MSC-Anchor links and the Anchor link to outside networks. Fiber cable is the transmission medium of choice for these link systems. Although microwave systems are sometimes used where fiber runs are not cost-effective, we assume the wireless infrastructure to be interconnected by fiber transmission capabilities. Fig. 6, 7 and 8 show the typical structure of link systems. Link systems can be generally classified as one of three cases:

- Case A is the single-fiber system with no backup (shown in Fig. 6).
- Case B has redundant fiber media backup. Redundant circuits are supposed to take different physical paths (shown in Fig. 7).
- Case C has fiber media, transceiver, and power backup, while transceivers are hot standby (shown in Fig. 8).

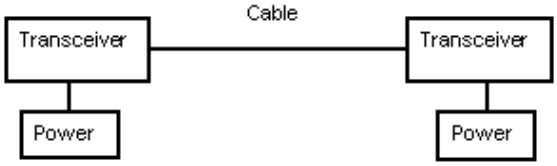


Figure 6. Unprotected Link

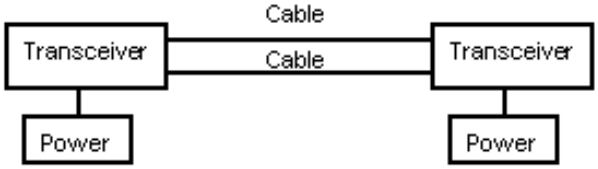


Figure 7. Partially Protected Link

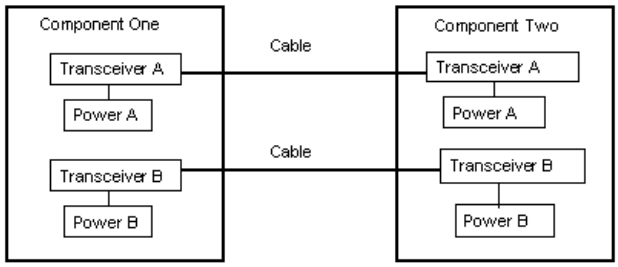


Figure 8. Fully Protected Link

From the multi-WIB infrastructure in Fig. 3, it is seen that although all links are important to a network's dependability, the reliability levels vary from one type link to the next. For example, the BS-BSC links are relatively less important than BSC-MSC links from the network operators' point of view. Similarly, BSC-MSC links are not as important as MSC-Anchor links. Each of the three link categories shown have different reliability or MTTF.

In Fawaz [22] fiber cable system reliability is discussed in detail. Statistics from Telecordia are referred to in that paper, where the authors came to three conclusions:

- The frequency of failure occurrence in optical network is not negligible.
- Cable cuts are the dominant failure scenario for long optical fiber networks.
- Power reliability is important in link reliability.

Table 4 shows their results. In this table, Failure In Time (FIT) is the average number of failures in 109 hours [22].

TABLE 4. Optical fiber and Transceiver Failure Rate [22]

Cable Cut rate FIT per 1000 miles	501142
Cable Cut rate per hour, 1000 miles	0.0005011
Cable Cut rate per year, 1000 miles	4.390
Cable Cut MTTF per 1000 miles (Yr)	0.228
Cable Cut MTTF per mile (Yr)	228
Transceiver Failure rate FIT	15178
Transceiver MTTF (Yr)	8.0
US Telecom Power failure rate per year	0.1252
Power MTTF (Yr)	8.0

The MTTF for links are different mainly because of the varying fiber length. If we assume that the hazard due to fiber cut will decrease linearly, the failure rate shown in Table 4 can be used in this work. For fiber links less than 10 miles, the transceiver and power systems become the dominant contribution to failure, rather than the fibers. This means that the total link system's MTTF is comparatively low for unprotected links. The MTTF of a parallel system with four components is about 1.6 times of the MTTF in the single system [9]. For instance, the MTTF of a 10-mile link without protection is given by:

$$MTTF_{10} = \frac{1}{\frac{2}{MTTF_{transceiver}} + \frac{1}{MTTF_{fiber}} + \frac{2}{MTTF_{power}}} \quad (5)$$

$$MTTF_{10} = \frac{1}{\frac{2}{8.0} + \frac{1}{22.78} + \frac{2}{8.0}} = 1.8 years \quad (6)$$

Likewise, the MTTF of a 10-mile, partially protected link is given:

$$MTTF_{10p} = \frac{1}{\frac{2}{MTTF_{transceiver}} + \frac{1}{1.6 * MTTF_{fiber}} + \frac{2}{MTTF_{power}}} \quad (7)$$

$$MTTF_{10p} = \frac{1}{\frac{2}{7.5} + \frac{1}{1.6 * 22.78} + \frac{2}{8.0}} = 1.9 years \quad (8)$$

Lastly a fully protected 10-mile link MTTF is given by:

$$MTTF_{10f} = 1.6 \times MTTF_{10} = 2.9 years \quad (9)$$

Table 5 shows MTTFs of fiber links under different protection schemes at different distances. From the calculation results, we can see that the MTTFs of fiber links at a distance between 10 to 20 miles are very similar.

TABLE 5. Optical Fiber Link MTTF

Link Length (Miles)	Optical Fiber MTTF (Years)	Unprotected MTTF (Years)	Partial Protect MTTF (Years)	Fully Protect. MTTF (Years)
1	222.79	3.9	3.8	6.3
5	45.56	1.9	2.0	3.1
10	22.78	1.8	1.9	2.9
20	11.39	1.7	1.8	2.7

According to the statistic data from US Census Bureau [23], the number of persons per square mile ranges from several hundred to over a thousand in metropolises such as Los Angeles, New York, Atlanta, and Phoenix. In this study, the following assumptions on fiber length and protection lead to the following MTTFs for links:

- Fiber link of BS-BSC is at 20 miles.
- Fiber link of BSC-MSC is at 5 miles level.
- Fiber link of MSC-Anchor switch is very short, less than 1 mile.
- Fiber link of Anchor switch-PSTN is very short, less than 1 mile.

A component's maintainability is represented by its MTTR. In order to understand the role that MTTR plays in dependability, three MTTR scenarios are used in the simulation: nominal, degraded, and enhanced. Nominal MTTR was obtained from [19]. The degraded MTTR was taken as three times the nominal MTTRs, excepting switches. Table 3 also lists the component MTTRs 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

IV. RESEARCH QUESTIONS

In this section, four major research questions are presented and discussed. Additionally, the assumptions made in addressing the research questions are listed.

Research Question 1: How will the number of impact epochs and their composition (number of concurrent component outages making up epochs) change as the network size, component reliability, and component maintainability change?

As customer demand increases in an area, the network size increases, and more components (equipment and links) are used. We expect that more component outages will occur as the network grows. The wireless infrastructure studied in this research, as shown in Fig. 3, indicates that the total

number of components failing is expected to scale with network size. We also expect impact epochs to grow along with network size—however, what is the relationship between the number of epochs and the network size? Will this number linearly scale as the network grows? Notice that impact epochs include both single and concurrent outage epochs, and as we count overlapping outages as one epoch, we expect the total number of impact epochs to grow nonlinearly.

Over time, how many impact epochs consist of more than one outage? The answer depends on several factors. First, more components mean more possible failures. So as the network grows bigger, the probability of simultaneous outages increases. This probability increases nonlinearly as the network size increases. The second factor is component MTTF. As component MTTF decreases, more component outages occur over a period of time. We expect multi-outage impact epochs to increase in a network as component MTTFs decrease. The third factor is component MTTR. If the repair time for a single outage increases, the probability for other outages happening during this repair interval increases. Thus we expect multi-outage impact epochs will increase as the component MTTRs increase. Network size, reliability, and maintainability interact in ways that make it difficult to predict either linear or non-linear behavior with regard to the number of impact epochs. This research investigates the relationship based on network size, reliability, and maintainability scenarios.

Research Question 2: What fraction of time is the network in a non-episodic state as network size, reliability, and maintainability change?

The percentage of time in one year that a network is in the quiescent state and non-quiescent state is insightful. The average quiescent availability is an important issue to network operators. The total non-episodic time is the sum of time that a network is in quiescent state over one year. It is expected that as the network size increases, the total time the network will be in a non-episodic state will decrease.

This question deals with how network size, component reliability, and component maintainability affect the total non-episodic time in a wireless network. More frequent failures and increasing repair times should decrease quiescent time. However, overlapping outages could increase quiescent time. How these factors combine to effect total quiescent time is not obvious.

Research Question 3: How will the dependability characteristics of impact epoch change with the network size, component reliability, and component maintainability?

Impact epochs have a number of characteristics such as MTTE, MTRE, and the peak number of impacted customers. As a system, a wireless network's MTTF is dependent upon all of its component's MTTFs. As the network size increases, the network component outages increase linearly.

MTTE is the mean time to epochs, instead of component failures within a wireless infrastructure. Because each epoch may be a single- or multi-component failure, the probability that an epoch includes more than one failure nonlinearly increases as the network size becomes bigger. So for MTTE, we expect it to decrease in a nonlinear fashion as the network expands.

The second attribute of impact epochs that is investigated in this work is MTRE. Due to increases in simultaneous component outages, MTRE increases as a network grows. How long MTRE lasts depends on how many impact epochs are multi-outage epochs. The higher the percentage of multi-outage epochs, the longer the MTRE will be. We expect a nonlinear growth on MTRE as the network becomes bigger.

The third impact attribute investigated in the research is Peak Customers Impacted (PCI). This factor shows how serious an impact epoch could be in the dimension of impact size. If we can find the distribution of PCI, we may be able to provide network operators the probability of an impact epoch impacting more than a set number of customers over a period of time. For example, we may calculate the probability of PCI exceeds 8000 customers. This could be valuable information to network operators.

Research Question 4: How will different thresholds help network operators filter impact epochs in a network?

Peak customers impacted (PCI) provides information of an epoch in only one dimension. Another perspective is one that considers size and duration of an epoch. In this research, the two-dimensional metric called WPLLH, discussed earlier, is also used to measure impact epoch. The WPLLH uses the wireless traffic profile developed earlier, rather than wire-line PLLH usage time factor.

Thresholds are powerful tools in network management because network operators usually prioritize their activities to respond to more important events in their networks. In this thesis, three different thresholds are investigated—5000 WPLLH, 10,000 WPLLH, and 15,000 WPLLH. The number of impact epochs over these thresholds is expected to grow as network size increases, or component reliability or component maintainability decreases. The number of epochs exceeding a threshold will change from one threshold to another. For example, the number over 15K WPLLH threshold may not change as fast as the number of epochs over a 5K WPLLH threshold. This is because epochs over 15K WPLLH rarely occur in smaller networks. This applies to different scenarios. For example, the number of epochs over 15K WPLLH is certainly less than that in a degraded reliability and maintainability scenario.

V. SIMULATION MODEL

Fig. 9 displays the input and output process of the simulation and the derived results, while Fig. 10 shows the

architecture simulated. Inputs for the simulation include all component MTTRs and MTTFs, wireless traffic profile, the network size, and an operational time of one year.

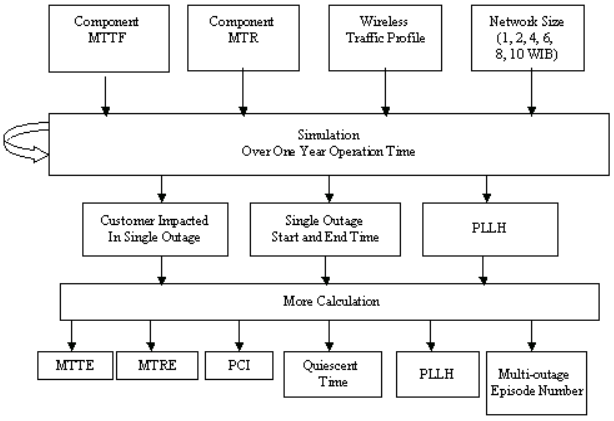


Figure 9. Process of Simulation and Results

Outputs from the program are network survivability as well as 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™.

To fully investigate effects of different levels of reliability and maintainability for different size networks (size determined by the number of WIBs), we investigate three scenarios: nominal, degraded maintainability as well as 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 6, showing four component outages, starting at 308.465 days into the year.

TABLE 6. 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. 11 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:

$$Q_t = \sum_{i=1}^{n} TTE_i = TotalSimulationTime - \sum_{i=1}^{n} TRE_i \qquad (10)$$

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 Fig. 11.

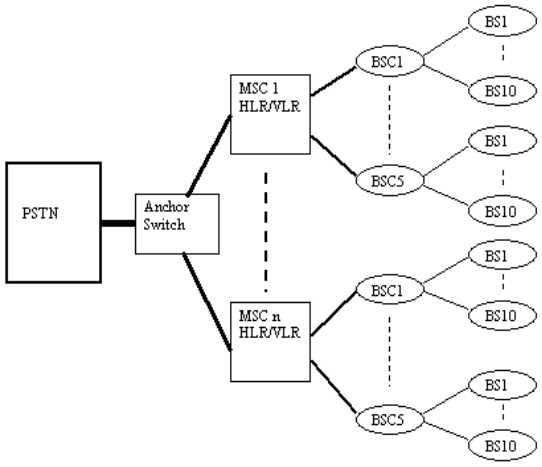


Figure 10. Scalable Network Size

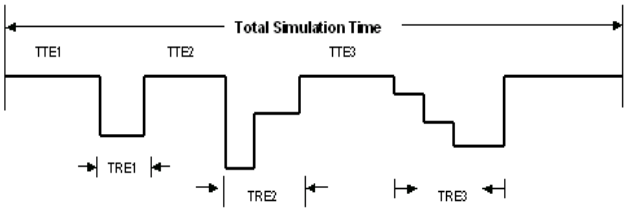


Figure 11. Relationship of TTE, TRE and Simulation Time

Likewise, we expect the MTTE (mean of all times to epochs TTE), MTTR (mean of all times to restore epochs TRE) and total simulation time to be:

$$MTTE = \frac{\sum_{i=1}^{n} TTE_i}{n} \qquad (11)$$

$$MTRE = \frac{\sum_{i=1}^{n} TRE_i}{n} \qquad (12)$$

$$Total\_Simulaiton\_Time = (MTTE + MTRE) \times n \qquad (13)$$

A. Further Model Verification

The discrete time event simulation model was written in VC++. All components/links in the WIB(s) are in service simultaneously. Component times to fail are exponential, while repairs times are lognormal. Simulated failure counts were exhaustively compared to expected component failure counts, while simulated repair times were compared to fitted lognormal repair distributions. The null hypotheses of "no difference" were accepted at high degrees of inference, using chi-squared test statistics.

Next, network survivability was checked for consistency against the three different scenarios investigated, and compared to analytic calculations. As explained earlier, survivability is defined as the fraction of WPLLH offered over a one year operation:

$$\text{Network Survivability} = 1 - \frac{WPLLH}{n \times TLH} \quad (14)$$

$$= 1 - \frac{WPLLH_A + WPLLH_{AL} + \sum_{m=1}^{n} (\sum_{i} WPLLH_i)}{n \times TLH} \quad (15)$$

where:

- *TLH* is Total Line Hour for one WIB (365×24× 100,000);
- *WPLLH* is Wireless Prime Lost Line Hour;
- *i* is the number of outages in the network;
- *n* is the number of WIB in the wireless infrastructure;
- $WPLLH_A$ is the prime lost line hours because of the anchor switch outage;
- $WPLLH_{AL}$ is the prime lost line hours because of the anchor link outage; and
- $WPLLH_i$ is the prime lost line hours for the i^{th} WIB in the network.

Based on the infrastructure used in this research, each WIB has the same structure, reliability, and maintainability levels, meaning that same-type component MTTF and MTTR are the same for each WIB in the architecture. So we may expect that each WIB will generate similar numbers of outages, outage repair times, and WPLLH. From the above equation, we can see that factors affecting network survivability are $WPLLH_A$ and $WPLLH_{AL}$. As any failure of the anchor link or anchor switch will impact the entire network no matter how many WIBs are in the infrastructure, we expect that the network survivability will stay relatively constant in each scenario, as survivability is the fraction of user hours available over a time period. However, nominal, degraded maintainability—as well as enhanced reliability and maintainability scenarios—will exhibit different network survivability levels. We expect that enhanced scenario to have the highest survivability because we expect the least outages. In contrast, the degraded maintainability scenario should have the lowest survivability because we expect the most outages.

The network survivability simulation results for each of the three scenarios is seen in Fig. 12. As expected, the enhanced network has the highest survivability and the degraded network the lowest. Also, for each scenario the network survivability remains constant for different network sizes, as expected. It also indicates that the simulation is verified with the new wireless traffic profile. In addition, the survivability by scenario and size were compared to analytic predictions and compared by chi-squared statistics.

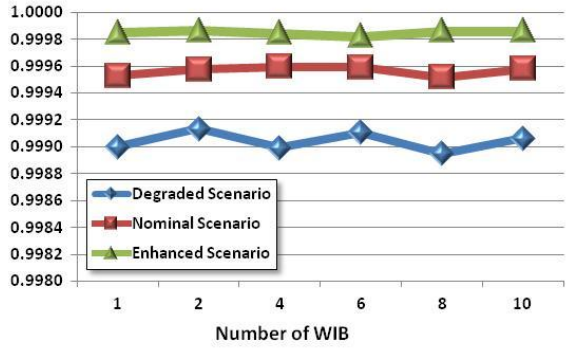


Figure 12. Network Survivability by Scenario

B. Assumptions and Limitations

The model is subject to the following assumptions and limitations:

- This work considers outages due to equipment and link failures ("components"), and does not focus on network disturbances due to traffic congestion.
- The wireless network under study is an infrastructure with an anchor switch as the gateway connecting to outside networks, such as the PSTN or other wireless networks. The anchor switch acts as the only interface to the outside world. All MSCs in this network will route their traffic with outside destinations to the anchor switch for further routing.
- An anchor switch is assumed to have the same dependability features as any other MSC in the network. The MSC has become very stable and reliable over many years of development.
- The network topology is a star topology, which is very popular in practice. The star topology distributes network functionalities geographically. It is assumed that there are no mesh topologies in the network.
- Structure and scale of all WIBs within the wireless infrastructure are the same.
- Nominal component MTTFs and MTTRs are based on published literature [10] and are not based on empirical data collected for this research.
- Component MTTF and MTTR are invariant over a one-year period. TTFs are exponentially distributed, consistent with a homogeneous Poisson process. TRs are lognormally distributed, consistent with long tail distributions to account for travel time.
- The impact on network dependability caused by anomalous propagation is not in the scope of this research as it relates to single outage.
- Fractional component failures are not considered in the research.
- Inter-WIB traffic is not modeled; however, impact

epochs in the research do include both incoming and outgoing communication loss.

- Optimal reliability and maintenance strategies are not addressed, as cost is not part of this research.

VI. RESULTS

Research Question 1: How will the number of impact epochs and their composition (number of concurrent component outages making up epochs) change as the network size, component reliability, and component maintainability change?

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. Fig. 13 illustrates the relationship between the total numbers of impact epochs at a different network size for each scenario over a one-year interval.

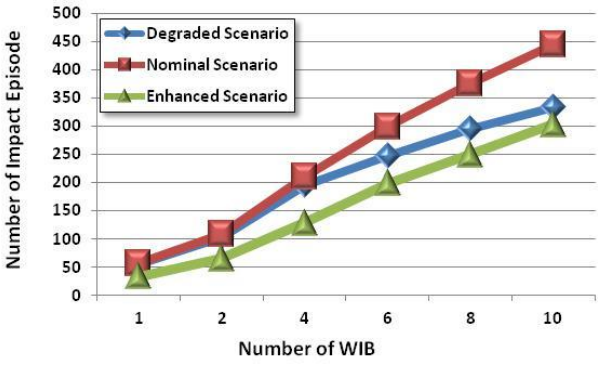


Figure 13. Total Number of Impact Epochs

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 that fewer components online at any instant can fail. As it turns out, the degraded case has less epochs, but more multi-outage epochs. Remember – a 1-WIB network serves 100,000 customers, while a 10-WIB network serves 1,000,000.

For larger networks that do not have enhancements in component reliability and maintainability, expanding a network presents challenges for network operators who must cope with impact epochs consisting of multiple outages that overlap in time. Repairing simultaneous outages is challenging in large networks especially because of geographic dispersion, which requires more maintenance staff, equipment, and vehicles. Component maintainability must be achieved even though there are simultaneous outages in the network. See Tables 7, 8, and 9 to see how the frequency of multi-outage epochs decreases as reliability and maintainability improve.

TABLE 7. Frequency of Multi-Outage Epochs: Degraded Scenario

No. Outages in Epoch	Number of WIB				
	2	**4**	**6**	**8**	**10**
1	65	125	189	234	281
2	1	4	9	15	20
3	0	0	0	1	1
4	0	0	0	0	0

TABLE 8. Frequency of Multi-Outage Epochs: Nominal Scenario

No. Outages in Epoch	**Number of WIB**				
	2	**4**	**6**	**8**	**10**
1	105	1	105	1	105
2	4	2	4	2	4
3	0	3	0	3	0
4	0	4	0	4	0
5	0	0	0	0	1
6	0	0	0	0	0
7	0	0	0	0	0

TABLE 9. Frequency of Multi-Outage Epochs: Enhanced Scenario

No. Outages in Epoch	**Number of WIB**				
	2	**4**	**6**	**8**	**10**
1	94	154	183	198	205
2	9	31	49	62	70
3	1	8	12	23	30
4	0	1	5	9	17
5	0	0	1	4	7
6	0	0	0	2	4
7	0	0	0	1	2
8	0	0	0	0	0

Research Question 2: What fraction of time is the network in a non-episodic state as network size, reliability, and maintainability change?

The simulated number of multi-outage epochs for each network size and scenario is displayed in Fig. 14. 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 10 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.

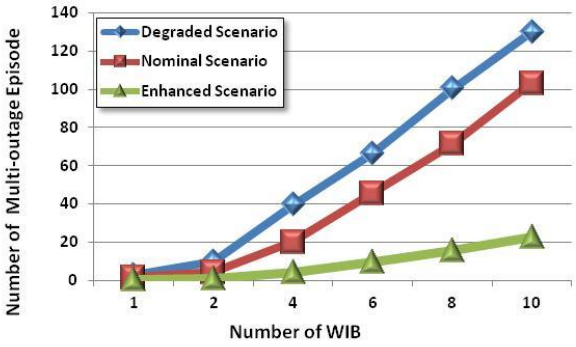


Figure 14. Multi-Outage Epoch Number

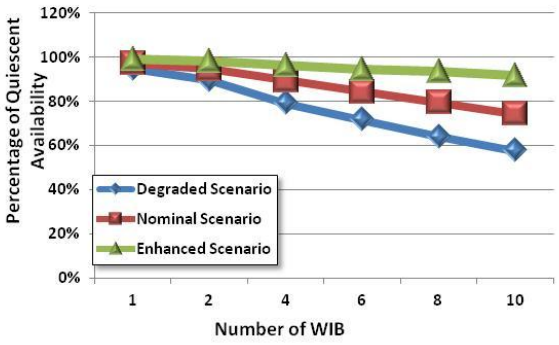


Figure 15. Percentage of Quiescent Availability

TABLE 10. Multi-Outage Impact Epoch Composition

# WIB	≥ 2 concurrent outages			≥ 3 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 the degraded and enhanced scenarios 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 scenarios. 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 Fig. 15. As the network expands, its quiescent availability decreases, almost linearly. In the degraded scenario, the total non-episodic time of a 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 a 30% improvement over the degraded scenario.

Research Question 3: How will the dependability characteristics of impact epochs change with the network size, component reliability, and component maintainability?

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. Fig. 15 shows the quiescent availability of a network in different scenarios. 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.

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, and is a measure of the network's maintainability. PCI and PLLH are subsequently 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.

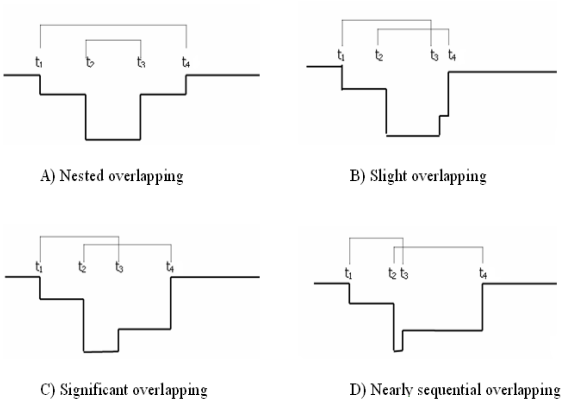


Figure 16. Different Overlapping Patterns

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 Fig. 16 A, B, C, D. Among these four patterns, pattern "A" does not increase TRE, since repair time of the second outage completely occurred within the repair time of the first outage (TRE in pattern "A" equals the MTTR 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. Fig. 17 illustrates the simulation output of the MTRE changes due to network size.

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 the nominal and enhanced scenarios. The MTRE of a 10-WIB network in the degraded scenario increased by approximately 28% (about 144 minutes) from the 1-WIB network, while a 10-WIB network in the enhanced scenario increased by only 5.4 minutes longer than the 1-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 [24] with a high degree of significance (p value less than 0.0001). This allowed easy calculation of probabilities of peak outages. Table 11 shows the exponential PCI means.

Table 12 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.

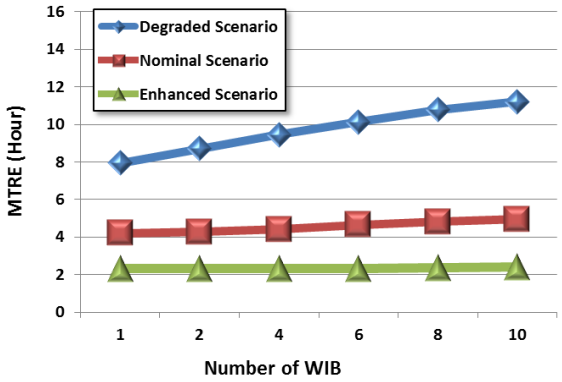


Figure 17. MTRE in Hours

TABLE 11. Mean PCI per Epoch

Scenario	Number of WIB				
	2	4	6	8	10
Degraded	2,932	3,296	3,509	4,407	4,549
Nominal	2,154	2,227	2,579	2,702	2,766
Enhanced	2,176	2,246	2,526	2,654	2,695

TABLE 12. 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%

C. WPLLH

Similarly, the distribution of WPLLH values for networks of different sizes and scenarios is illustrated in Table 13. These results can predict the probability of PLLH over a threshold for a given time period.

TABLE 13. 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

Research Question 4: How will different thresholds help network operators filter impact epochs in a network?

The chance of the PCI and the WPLLH 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 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 could mean 5,000 customers are impacted for one hour, or it could signify that 10,000 customers are impacted for half an hour. Fig. 18 indicates the relationship between the numbers of impact epochs versus different thresholds for the degraded scenario.

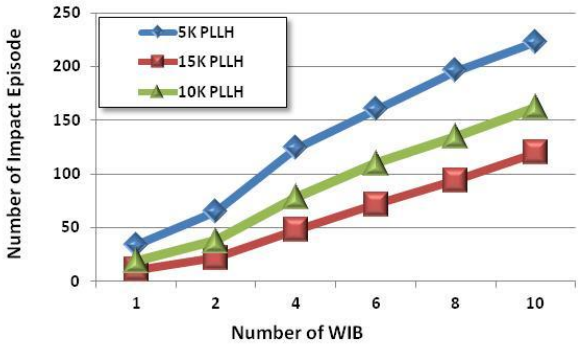


Figure 18. Number of Impact Epochs with Filters in Degraded Scenario

The growth rate of impact epochs over 5K WPLLH in all three scenarios increases 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 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, and too many less significant outages are seen.

VII. CONCLUSION

New dependability metrics have been developed here to investigate concurrent multiple outage epochs. Results indicate 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 a 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:

- Defining the impact epoch as a new way to evaluate wireless network infrastructure's dependability.
- Developing new metrics for analyzing RAMS for large networks (MTTE, MTRE, Quiescent Availability, PCI and WPLLH).
- Development of empirically derived wireless traffic profiles to determine the 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 number 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 both component reliability and 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 an acceptable minimum.

REFERENCES

[1] Andrew Snow, Gary Weckman, and Yachuan Chen, "Multi-Episodic Dependability Assessments for Large-Scale Networks", ICN 2011: The Tenth International Conference on Networks, pp. 441 to 448, IARIA, St. Maarten, The Netherlands Antilles, January 23-28, 2011, ISBN: 978-1-61208-113-7.

[2] S Hariri, S., et al, "Impact analysis of faults and attacks in large-scale networks", Security & Privacy, IEEE Sept.-Oct. 2003, Volume: 1, Issue: 5, Page(s): 49 – 54.

[3] B. Bassiri and S. Heydari, "Network Survivability in Large-Scale Regional Failure Scenarios", Proceedings C3S2E '09 Proceedings of the 2nd Canadian Conference on Computer Science and Software Engineering, Pages 83-87, ISBN: 978-1-60558-401-0, ACM New York, NY, USA 2009.

[4] J. Appleyard, "Outage management portal leveraging back-end resources to create a role and user tailored front-end interface for coordinating outage responses", Patent US 8,171,415 B2, May 1, 2012.

[5] Avižienis, A., Laprie, J.-C., Randell, B., & Landwehr, C. (2004). Basic Concepts and Taxonomy of Dependable and Secure Computing. IEEE Transactions on Dependable and Secure Computing, 1(1), (pp. 11–33).

[6] J. Knight, E. Strunk, and K. Sullivan, "Towards a rigorous definition of information system availability". Proceedings of the DARPA Information Survivability Conference and Exposition, IEEE, 2003.

[7] A. Birolini, Reliability *Engineering: Theory and Practice*, Six edition, Springer, 2010.

[8] M. Ayers, *Telecommunications System Reliability Engineering, Theory, and Practice*, IEEE Press Series on Network Management, John Wiley and Sons, 2012..

[9] A. Snow, "A survivability metric for telecommunications: insights and shortcomings", 1998 Information Survivability Workshop – ISW'98 IEEE Computer Society, FL, October, 1998, pp.135-138.

[10] A. P. Snow, "Assessing pain below a regulatory outage reporting threshold", Telecommunications Policy", Vol. 28/7-8, 2004, pp. 523-536.

[11] Andrew P. Snow, "The failure of a regulatory threshold and a carrier standard in recognizing significant communication loss", TPRC 2003, November 2003.

[12] Carol Y. Carver and Andrew P. Snow, "Assessing the impact of a large-scale telecommunications outage", Proceedings of the 7th International Conference on Telecommunication Systems, March, 483-489, 1999.

[13] H. Cankaya, A. Lardies, and G. Ester, "Availability-aware analysis and evaluation of mesh and ring architectures for long-haul networks", Applied Telecommunications Symposium, pp. 116 – 121, 2004.

[14] CTIA Wireless Quick Facts, retrieved from www.ctia.org in September 2012.

[15] U. Varshney and A. Malloy, "Multilevel fault-tolerance for designing dependable wireless networks", Proceedings of the 36th Annual Hawaii International Conference on System Sciences, HICCS 03, IEEE, Jan. 2003.

[16] FCC , Report and Order and Future Notice of Proposed Rule Making, retrieved from http://www.fcc.org in August, 2005.

[17] FCC, Public Safety and Homeland Security Bureau, www.fcc.gov/911/enhanced/, retrieved September 2012.

[18] K.. Du, and M. Swamy, *Wireless Communication Systems: From RF Subsystems to 4G Enabling Technologies*, Cambridge University Press, 2010.

[19] A. Snow, U. Varshney, and A Malloy, "Reliability and survivability of wireless and mobile networks"., IEEE Computer Magazine, July, 2000, pp. 49-55.

[20] Committee T1, T1A1.2 Working Group on Network Survivability Performance, "Enhanced Network Survivability Performance", Technical Report, November 2000.

[21] M. Albaghdadi and K. Razvi, "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, 2005.

[22] W. Fawaz, F. Martignon, K. Chen, and G. Pujolle, "A novel protection scheme for quality of service aware WDM networks", IEEE, 0-7803-8939-5, 2005.

[23] U.S. Census Bureau reports, retrieved from http://quickfacts.census.gov/qfd/states/ in September, 2012.

[24] BestFit 4.5, Palisade Corporation.

Integrating Future Communication Technologies for the Downstream Component of Public Warning Systems

Michelle Wetterwald, Christian Bonnet,
EURECOM, Sophia Antipolis, France
michelle.wetterwald@eurecom.fr
christian.bonnet@eurecom.fr

Sebastien Grazzini
Eutelsat, Paris, France
sgrazzini@eutelsat.fr

Daniel Camara
INRIA, Sophia Antipolis, France
daniel.camara@inria.fr

Xavier Ladjointe, Jean-Louis Fondere
Thales Alenia Space, Cannes, France
xavier.ladjointe@thalesaleniaspace.com
jean-louis.fondere@thalesaleniaspace.com

Jérome Fenwick
Groupe SYNOX, BALMA, France
jfenwick@groupe-synox.com

***Abstract*— Natural disasters have often made the headlines in the past years. As a consequence, many actions have been started by the public authorities to reduce the damages and the number of casualties. In that objective, the French project RATCOM aimed at developing an alert system in case of coastal tsunami due to underwater landslides. Its downstream component combines reliable and efficient communication systems to relay the alert. In parallel to the integration of the existing technologies in the project demonstrator, a survey analysis has been performed to identify the communications technologies and networks, which are in preparation but not yet operational, and which will increase the efficiency and quantity of individuals reachable by the future population alert networks. Each of these technologies is not sufficient by itself, but their combination improves drastically the efficiency of the alerting global system. This paper presents the RATCOM architecture, focusing mainly on its downstream component. For each candidate technology, it analyses how it can satisfy the requirements and improve the efficiency of the public alerting system. The final demonstration of the project is also described, as it assesses the feasibility of the system and how the overall impact on the alert dissemination is improved by the design of this new architecture.**

Keywords - tsunamis; alerting; public warning system; broadcasting networks.

I. INTRODUCTION

Natural disasters and the thousands of casualties they usually cause raise a major concern at public authorities' level. A milestone event in this field was the Indian Ocean tsunami that happened in December 2004. This event raised the question of how to improve the protection of the population and prevent so many deaths. In fact, the main answer relies in the fast distribution of the information: information about the best behaviour to adopt in case of a disaster, and more importantly, information about the imminent arrival of a disaster.

The South East part of the French Mediterranean coast has been identified by the experts as the potential location for small-sized tsunamis. These could be caused by major landslides in the underwater area, few kilometres away from the coast. One of these tsunamis occurred in 1979 in front of Nice and made several million euros' worth of damage. As a prevention tool, the RATCOM project [1], started in 2009 and ended in June 2011, aimed at developing and confirming the feasibility of an evolved alerting system towards the public safety professionals on one hand and the citizens on the other hand. The project has been organized around two major components: the upstream component and the downstream component. The upstream component is responsible to monitor the events occurring at the sea and report the risk level to a Control Centre. The Control Centre then makes the decision to generate an alert and forwards it to the downstream component, which is responsible to disseminate the warning within the shortest time frame possible. Mainly the downstream component is addressed by this paper.

The best-known method to broadcast alert messages is by triggering the operation of alert sirens. However, more modern technologies exist nowadays that can help reaching a larger quantity of people. The RATCOM downstream component aims at identifying and setup a network combining these technologies into a single framework. Some of these technologies are currently operational and have been included in the final project demonstration. To complete this setup, an additional survey paper activity has been conducted to identify other technologies that are not ready to be

included in the demonstrator, but may become relevant to this warning system in the future. Their suitability to be included in the project downstream component has been analysed. The final objective of this study, which is reported in this paper, is to draw up an inventory of the technologies and networks that are not yet operational, but are relevant to be used in the context of a future public warning system. Understanding what these technologies can bring and how they can be included in the system architecture that combines them as much as possible is an important step for the definition of future systems. This study has provided the core of [2], which is extended here to provide more details on the RATCOM objective and architecture, the prospective technologies, and a presentation of the feasibility demonstration that was performed at the end of the project.

The rest of the paper is organized as follows. Section II describes the RATCOM architecture and its expected impact on the alert dissemination. The third section considers systems of communication close to their deployment phase, with a probable delay of less than three years, and having the ability to be connected to a warning system in the medium term. Derived from digital broadcast systems, the DVB-SH (Digital Video Broadcasting for Satellite Handheld) uses the coverage capabilities of satellite networks. Satellites offer also the possibility to provide redundant connections and improve the strength of the whole system. WiMAX (Worldwide Interoperability for Microwave Access) is a new technology, which can provide service to larger areas than Wi-Fi (or IEEE 802.11), whose concept is somewhat similar. The new capabilities and possibilities of connecting current and upcoming mobile cellular networks are discussed. In the third part are presented prospective technologies that are currently being defined and standardized, but which will be effectively operational in a period longer than five years. They are essentially the Public Warning System integrated in mobile phone networks, broadcast technology in these global networks and Vehicular Networks. Finally, we draw our conclusion to this study in the last section.

II. THE ALERT SYSTEM ARCHITECTURE

Following the tsunami that occurred in December 2004 in the Indian Ocean, the French government ordered a risk analysis covering the French coast which highlighted that the south-eastern part ran a small risk, possibly triggered by large underwater landslides. Willing to develop activity and competence in this area, the government decided to launch a project aiming at the feasibility study for an end-to-end system addressing that challenge, i.e., short range local tsunamis, somehow different from the earthquake and tsunami monitoring systems deployed in the Pacific Ocean for example, which address long range threats.

Providing an end-to-end architecture, from seismic or pressure sensors to public warning, the RATCOM project has been organized according to two main components. The upstream component monitors the threat, while the downstream component disseminates the alert. Both are linked through a terrestrial or Alert Elaboration Centre, also called Control Centre. These components are pictured in Figure 1. The upward component has the objective to deploy sensors in the sea or on the coastal ground, and to provide an enhanced automated aggregation and analysis of their outputs, in order to help making the alert decision, while eliminating the risk of false alarm.

This processing is performed in the Control Centre and, when relevant, launches the alert towards two different groups: firstly the local and administrative authorities through a highly secured specialized network, secondly the concerned population (mentioned as citizens in the figure), through a public warning network organized around a centralized and intranet-like network called SecuNet. Even though it would have been interesting to describe the whole chain in details, this paper targets the study that was performed to enhance the public warning system with modern communication technologies, in order to extend the impact of the alert on the largest number of people possible and comply with a requirement of an alert dispatched in a few seconds timeframe.

Figure 1. RATCOM project components

When studying existing alerting systems, including in France, a varied set of mechanisms can be found, addressing diverse types of population. Moreover, they are usually not correlated, lacking thus the benefits of an integrated system. Finally, they often rely on old technologies like sirens or direct voice communications, obviously not benefiting from the more recent technologies brought by satellites, mobile cellular networks or, in the future, vehicular communications. The impact of the integrated system highly depends on the combination of technologies used for the alerting system. It varies also according to the density and location of the population in the target area. The same beach in August or in November would not host the same number of people. It's even more difficult to evaluate if we include the first ten meters of coast where people would or would not be swimming. Some parts of the coast offer beaches; others are sided by roads, where people are driving their cars, or railways with frequent trains. The objective and novelty of the RATCOM downstream component is thus to federate a large amount of communication and warning system in order to become more adaptive to the conditions under which the alert has to be broadcasted. Next section will analyse how

the state of the art technologies are suited to be federated under the SecuNet umbrella, while Section IV will evaluate the relevance of future upcoming technologies.

III. USING STATE OF THE ART TECHNOLOGIES

In this section, some technologies that are still in their early phase of deployment or will be in the next two or three years are described. The section focuses only on the technologies that seem to be relevant for the broadcast of warning messages to the public. These technologies (satellite systems, WiMAX and cellular systems) have the ability to quickly reach a greater proportion of the population, including people on the move. They constitute a part of the population who could not be informed by more traditional methods such as the television. For each of the technologies is presented a fast description then an analysis of why it is relevant for the public warning and for the interconnection with our alert distribution system is performed.

A. Satellite systems

Satellite systems can be used in two different manners: first manner consists in broadcasting directly information to handheld devices. This is the DVB-SH technology. The second manner consists in strengthening the whole system by operating connections redundantly with terrestrial links, which may be at risk.

The DVB-SH [3] is a standard derived from the DVB-H (Digital Video Broadcast-Handheld) standard to distribute broadcast video, audio and data to mobile devices such as mobile phones. Mobile TV (television) is definitely set to become the next major media market of tomorrow. The publication in November 2004 of the DVB-H standard, seen by analysts as a possible solution for providing mobile television, was the starting point of a series of work on this new mode of television programs consumption. While DVB-H is designed primarily for use in the UHF terrestrial broadcasting only, the DVB-SH tries to exploit the S band, as shown in Figure 2, where there are opportunities for Mobile Satellite Services (MSS). Thus, this standard, created specifically for distributing content in mobility situation, makes a major innovation in the satellite telecommunications world: it enables the addition of a network of terrestrial repeaters, called CGC (Complementary Ground Component) to complement the satellite coverage. This is displayed with the terrestrial repeater in the middle of the figure.

One of the major problems in terms of warning systems is to quickly reach a large number of people, whether they are in a mobility situation or not and, if possible, at a reduced cost. The DVB-SH broadcast network meets these criteria through the variety of devices able to receive the signal (mobile phones, vehicular terminals, etc.) as well as through the possibility of sharing the same flow between a large number of people via the satellite. Accordingly, it becomes quite interesting to interface our alerting network and demonstrate the potential offered by hybrid broadcast architecture. Three warning systems are considered in the framework of this study: first, the broadcast of video / audio warning on TV / Radio mobile satellite devices, second, the broadcast of a detailed report about the alert to the TV / Radio mobile satellite devices for interested people and finally the triggering via the satellite of fully autonomous and easily installable alerting peripherals (e.g., on beaches).

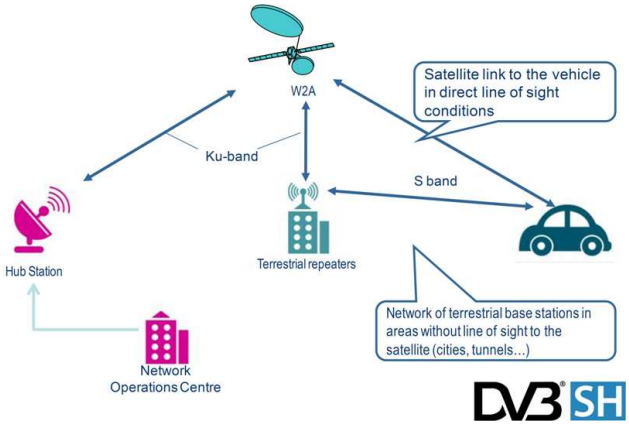


Figure 2. DVB-SH broadcast network architecture

The first two cases are closely related; they actually consist in stopping the Radio / TV programs to replace them with the tsunami warning. The procedure is very similar to what exists for the abduction alerts on TV but would be applied to mobile TV. The major innovation lies in the fact that, simultaneously with the program stop, an alert is sent as data traffic and the user can view this bulletin with the same device. This bulletin can be updated to indicate, for example, the end of the alert. The third case is the satellite triggering of alerting fully autonomous and of easy maintenance devices. Indeed, with DVB-SH, it is possible to receive the signal with a small omnidirectional antenna and one can imagine devices (sirens, billboards, etc.) independent of terrestrial communications networks that can be triggered remotely via a satellite signal. This new type of installation would benefit from reduced costs because no wired connection would have to be planned and its assembly and disassembly in urban areas would be simplified. The positioning of the devices would be defined only by taking into account the risk factor and not the availability of a terrestrial network. This freedom enables an improvement of the efficiency of the devices. Moreover, such a warning system would benefit from a complete independence from terrestrial communications networks, which can be damaged by natural disasters.

As the second manner to use the satellite technology, the Ku band connection systems or VSAT (Very Small Aperture Terminal) serve redundant network nodes or quickly connect fixed subscribers or isolated alert networks. This is illustrated in Figure 3. These systems make use of satellite dishes with a diameter less than 3 meters and terminals (or modem) that allow bidirectional communications. They provide the following intrinsic advantages: a minimum ground infrastructure, an immediate area covering several alert networks from one or several countries at the same time and a simple and rapid deployment. With a satellite link, it is possible to connect either a comprehensive warning system,

in which case we preferably connect via the satellite the control node responsible for the warning broadcast on this network, or a specific node of the warning network, such as a siren, a VMS (Variable-Message Sign) or any other equipment that would require redundancy or that just needs to be connected to the network. Such a node can be a warning system sharing the same satellite link or a single important subscriber connected to the satellite endpoint.

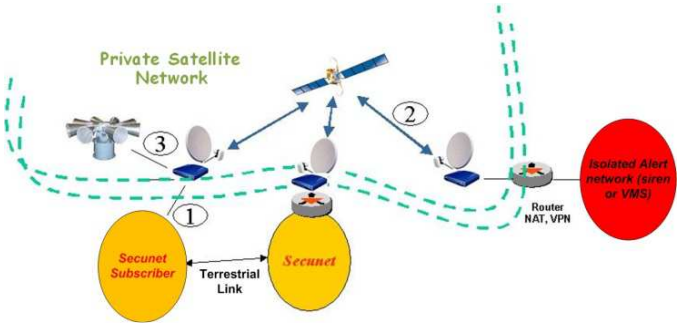


Figure 3. Ku Band Satellite Connection Systems

The choice of the satellite as the transmission system component on the downstream component is justified by the desire, first, to avoid congestion or interruption of the terrestrial networks, which can become harmful in case of a tsunami, and, secondly, to be able to quickly connect a warning system or a single isolated but important subscriber. In this case, the satellite will thus be used for the redundancy of critical network nodes (connected to a warning system or a subscriber of critical importance in the decision process), to connect the system to an existing warning network, or just to quickly connect a siren or isolated warning sign.

B. WiMAX networks

The WiMAX technology is standardized by the IEEE (Institute of Electrical and Electronics Engineers) under IEEE 802.16 and addresses several objectives: fixed mobile convergence, higher flow rates, compliance with quality of service constraints, etc. Compared with the architecture of conventional cellular systems such as EDGE (Enhanced Data for [Global System for Mobile communications] GSM Evolution) or UMTS (Universal Mobile Telecommunications System), the architecture of a WiMAX network is based on components that are intended to remain close to the Internet standards, as pictured in Figure 4.

The standard provisions various types of communications. For point to point transmission, it aims to link transmission points separated by a few dozen kilometres for the multiplexing of IP traffic with the support of differentiation and service guarantee. This type of application is similar to radio-relay transmission while it provides the spectral efficiency and intelligent management of IP traffic. It comes in support of network deployments that would not be economically viable if done in wire line technologies. The systems for point to multipoint transmissions without mobility provide the Internet traffic from a connection point of the wired network to a group of buildings or homes through the radio interface. User devices within the buildings are basically PCs (Personal Computers) that receive a service equivalent to an ADSL (Asymmetric Digital Subscriber Line) access. This standard thus targets to address the so-called "white areas" in which a typical deployment of ADSL based on a wired infrastructure would be too expensive to setup. In the point to multipoint transmission with mobility version, the WiMAX radio signal terminates directly on the terminal of the final user. This system can accommodate the wireless ADSL users, but also PC terminals (usually laptops) for a mobile Internet access.

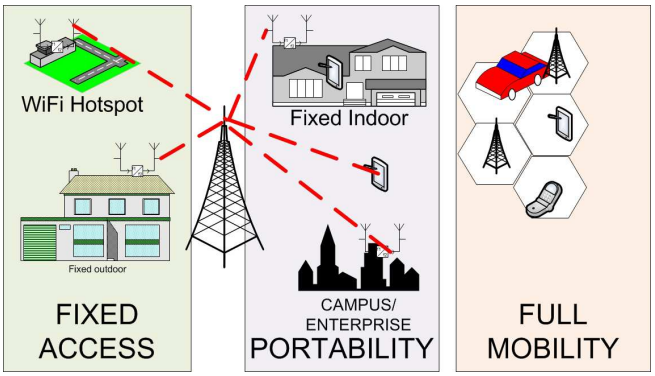


Figure 4. WiMAX Mobile Environment

The WiMAX offers a continuous connection for the transfer of IP packets. Accordingly, it can support any type of warning system based on data transmission. An interesting feature is its ability to support a Broadcast / Multicast mode called MCBCS (Multicast and Broadcast Services). In the same perspective as the 3GPP MBMS (Multimedia Broadcast/Multicast Service) technology, the WiMAX plans to provide broadcast services in geographical areas managed by the system. In a Multi-BS (MBS) system, several BSs located in the same geographical area, called MBS_ZONE, can transmit the same broadcast / multicast messages simultaneously on a single radio channel. It should be noted that a BS may belong to several MBS_ZONEs. A mobile terminal that registers for an MBS service can receive information from all the BSs of the MBS_ZONE without having to register with a specific BS of the area. In addition, it can receive the MBS signals from several BSs (Base Stations) simultaneously for an improved reception quality. This broadcast service enables the usage of the WiMAX technology as a potential support for public alerting messages.

C. 3G and LTE cellular networks

The CBS (Cell Broadcast Service) technology allows sending through the GSM network one or more small messages to all the mobile phones located within a specific area covered by one or several broadcasting Base Transceiver Stations (BTS), as pictured in Figure 5. The information can be broadcast over several channels, possibly one per language used for the broadcast message. The user must first select the channels to which he wants to subscribe.

This technology allows disseminating a mass message without network performance problems. However, setting the terminal requires an adequate communication plan to the population associated with a technical support team able to assume the setup on heterogeneous consumer devices, in the best case when they are compatible, since the CBS feature has been removed from many terminals in favour of more vending features. In any case, this technology has been selected by the standards to carry the messages of the Public Warning System (PWS), as will be explained in Section IIIB.

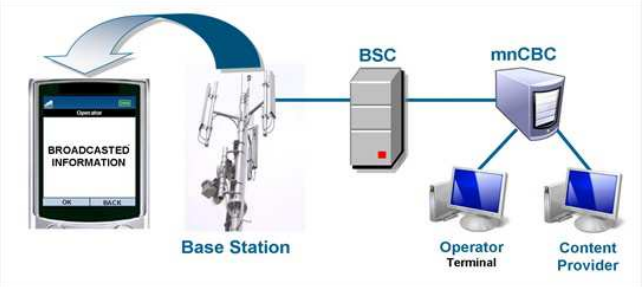


Figure 5. Using 3G and LTE (4G) Cellular Networks

The LTE (Long Term Evolution) is a project led by the 3GPP standards body for the publication of the technical standards of the future fourth generation mobile telephony. It enables data transfer at very high speed, with a longer range, a higher number of calls per cell and lower latency. For the operators, the LTE involves changing the core network and the radio transmitting stations. New compliant mobile terminals must also be developed. Considering the limitations of the current solutions in terms of deployment and performance, the LTE generation allows, with continuous connections, to be able to alert all the terminals almost simultaneously in a specific area, using dedicated short messages. The question of the penetration rate of terminals with 4G subscriptions is an important element in the relevance of the solution for an alerting system. The number of users accessing the 3G services has been increasing sharply since the latest developments of devices such as the iPhone, Android, BlackBerry or Windows phones and the commercialization of unlimited flat rate packages. The population currently reached with 3G mobile subscriptions will probably evolve to the upcoming 4G systems rapidly due to the effect of device renewal.

IV. USING ENHANCED UPCOMING TECHNOLOGIES

This analysis has been completed with a prospective study of networks currently in the phase of definition and standardization, and which are of interest for the future population warning systems. In a first step are introduced the future vehicular networks, which deployment is planned for the second half of the decade. The advantage of such networks is that, in addition to being able to reach the drivers, they operate in a cooperative mode. As a result, these networks are resilient to the possible destruction of the communications infrastructure. In a second step are presented the future developments of broadcast technology for mobile cellular networks (CBS and MBMS) and their integration in terms of standards into warning systems. The presented techniques were initially developed for a tsunami warning network in Japan and subsequently generalized to a more comprehensive Public Warning System (PWS). The CBS technology is used here again. This standard is part of the GSM, UMTS-3G and future LTE operational standards. Its advantage is that it allows the global broadcast of short messages (SMS-type) and thus overcomes the limitations due to network overload when targeting a large population. It also contains features that allow to "wake up" idle mobile phones and select the geographical coverage for the broadcast, making it particularly suitable for a connection to a global alerting system. However, it is somehow questioned since its deployment differs according to operators and countries.

A. Vehicular Communications

This new mode of communication from vehicle to vehicle is based on the new standards for Intelligent Transport Systems (ITS). Here we introduce the ETSI TS 102 636-3 standard [4], which specifies the GeoNetworking operation in the ITS environment. The most interesting feature of this standard is the definition of a set of methods to distribute, and route messages in specific geographical areas. For example, in the advent of an emergency situation, the message is sent to the vehicles concerned by this emergency in the destination area. It would, in this way, reach only the concerned vehicles, not disturbing drivers outside the target region. The communication among entities may be between Vehicle to Vehicle (V2V), Infrastructure to Vehicle (I2V), Vehicle to Infrastructure (V2I), Infrastructure to Infrastructure (I2I) and all the concatenation of these basic scenarios. Vehicular communication is becoming a huge research area and one of the most crucial aspects into this new research field is the data forwarding problem. Data forwarding is related to how to transmit a packet from one node to another, trying to reach the destination. Usually, in the context of vehicular networks, nodes forward data through geocast, where the position of the nodes defines the way the data will be transferred. There are basically three types of data forwarding schemes: geographical unicast, geographical broadcast and topologically scoped broadcast. Figure 6 (a) shows an example of geographical unicast, where multi-hop data transfer is used to connect the origin and the destination. Only one copy of the message is present at each time in the network. In geographical broadcast, Figure 6 (b), the message is distributed by unicast until a delimited region, where the nodes rebroadcast the message using flooding. In topologically scoped broadcast, Figure 6 (c), the nodes rebroadcast the messages for a predefined number of hops from the origin. Propositions and techniques, linked to data forwarding, range from the use of torrent-based communications to propagate messages, to the seamless connection to the network and the use of Delay Tolerant Networks (DTNs).

As a work on seamless connectivity, we can highlight BATMAN [5], a distance vector based routing protocol that performs channel selection between vehicular and roadside

mesh. On BATMAN, each node has its own forwarding strategy to find the best next hop and reach the destination. The proposed solution shows to be more efficient than some of the most popular routing protocols for mesh and ad hoc networks.

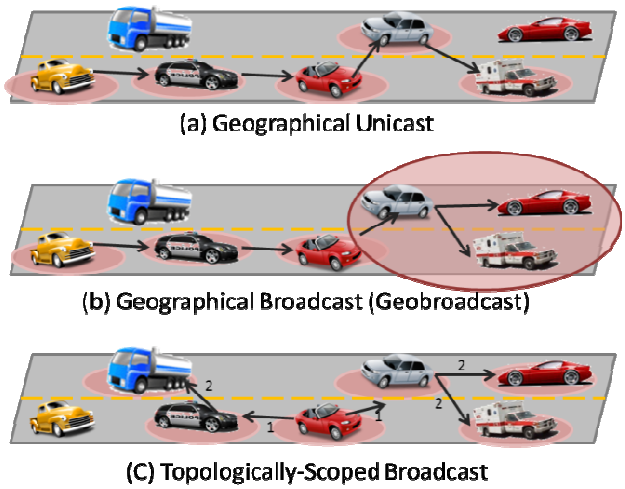


Figure 6. Types of V2V communication, (a) geographical unicast (b) Geobroadcast (c) Topologically-scoped broadcast

Some additional research has been conducted on the propagation of public safety warning messages using GeoBroadcast and Delay Tolerant Networks techniques [6]. The main purpose of such work is to increase the coverage of the existing network to reach more people in a faster way. People in vehicles usually do not watch TV and may not be listening to the radio. In the future, cars will be equipped with devices helping to increase road safety that will be constantly active to provide the drivers with the information about the road conditions. The work described in [6] proposes that the vehicles act as virtual Roadside Units (vRSUs) and help on the spreading of the warning messages in case of an emergency. The intention is to decrease the "last mile" information access problem. The evaluations show that the mechanism is robust and efficient even over different disaster scenarios. Thus the use of vRSUs is an effective way to distribute warning messages to vehicles in a region. One of the greatest advantages of this kind of epidemic approach is its efficiency. Remembering that the target scenario for this work is the propagation of public safety warning messages, i.e., extremely important data. vRSU, even considering disaster scenarios can redistribute a warning message to all nodes into an area of 15x9 kilometres in about six to seven minutes.

B. *Future Cellular Technologies*

Some new technologies and actions have recently been introduced in the 3GPP standardization for cellular systems which are relevant to public warning systems. The first part describes the two candidate technologies that can comply with the broadcasting requirements in case of a major event. Both technologies offer a global broadcast capability, which means that a message is sent only once and received at once by all the target terminals.

The CBS, which has already been pointed out above (Section IIC) as a potential existing technology, has been part of the standards since the early GSM, even if not always deployed by operators, so it is technically compliant with all the existing enabled mobiles in the market. It permits to broadcast unacknowledged messages to all the receivers within some particular defined geographical areas known as cell broadcast areas. A CBS page is comprised of 93 characters and up to fifteen pages may be concatenated to form a message. Messages are broadcast cyclically at a frequency and for a duration agreed with the information provider. Mobiles can selectively display only the messages chosen by the Mobile user. In addition, a message that has been formerly successfully received is not displayed a second time. The second technology, the MBMS is an enhancement of the 3G systems which provides a point-to-multipoint capability for Broadcast and Multicast Services [7], allowing resources to be shared in the network. Figure 7 shows the network reference model with the infrastructure design of the MBMS, as defined in the 3GPP standards for cellular network. Since it is more recent than CBS, it has more constraints, but it also brings the capability to disseminate multimedia information (video, audio, pictures) in addition to the text messages. As the LTE is enhancing the capacity and efficiency of the cellular networks, the MBMS is evolving and adapted to benefit from these improvements.

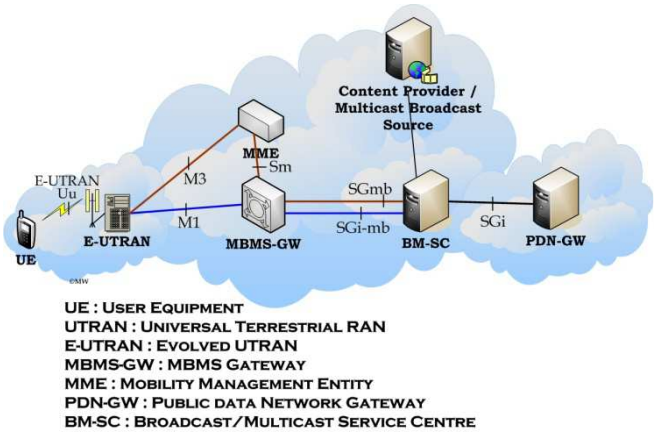


Figure 7. MBMS model in the LTE system

Public notification warnings have recently been implemented by the 3GPP standardization groups. Japan launched the first step with the ETWS (Earthquake and Tsunami Warning System), delivering Warning Notifications specific to Earthquake and Tsunami simultaneously to many mobile users located in Warning Notification Areas, typically a distribution of cells, who should evacuate from an approaching Earthquake or Tsunami. The architecture and notification hierarchy of the ETWS is shown in Figure 8. An ETWS warning may be required in a very urgent timeframe (down to 4 seconds for the primary notification or initial

alert) and is characterized by the capability to provide a very short notification period. A secondary notification can be delivered afterwards, carrying a larger amount of information such as text, audio or graphics to instruct what to do and where to get help, or a valid route from present position to an evacuation site. In a further release, this system has been generalized into the PWS (Public Warning System) [8], which targets worldwide objective, including the CMAS (Commercial Mobile Alert System) in the USA or the support of European requirements. The minimum functionalities to be supported by warning providers are activation of the notification delivery, its update and its cancellation. This notification must be delivered without any user interaction, even if it targets a terminal in sleeping mode. On the contrary, a manual action is mandatory to suppress the message, increasing the potential impact of the method.

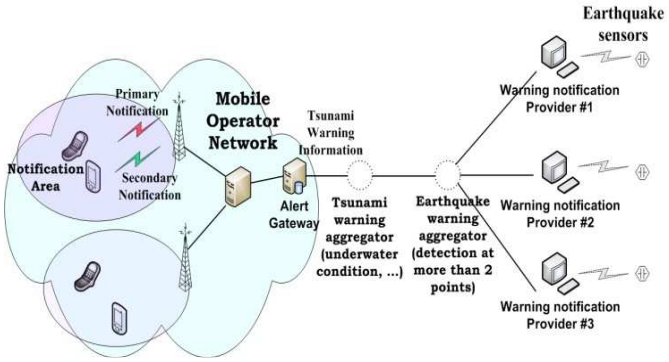


Figure 8. Global view of the Earthquake and Tsunami Warning System

Some early technical studies considered both some variants of the CBS and MBMS broadcasting technologies for the PWS. Since the CBS is more mature from a standardization point of view, it is the solution that has been adopted. However, because the MBMS will be part of the future EPS (Evolved Packet Systems), which will replace the current mobile networks, and can convey larger amount of data, it is still an interesting candidate to support future alerting systems. One of its drawbacks, though, is that it lacks the geo-localization feature of the CBS system. An enhanced system has been proposed in [9] to extend the MBMS by developing cross-layer cooperation where the networking protocol and the cellular system collaborate to improve the efficiency of the geographical radio coverage. It enables a more precise and efficient delivery of the broadcast information, taking advantage of the comprehensive knowledge of the infrastructure and network topology by the mobile operator. Only the base stations located in the target zone participate in the distribution of the message, as shown in Figure 9.

Users located outside of their coverage do not have to filter out the un-necessary information, increasing the efficiency of the system.

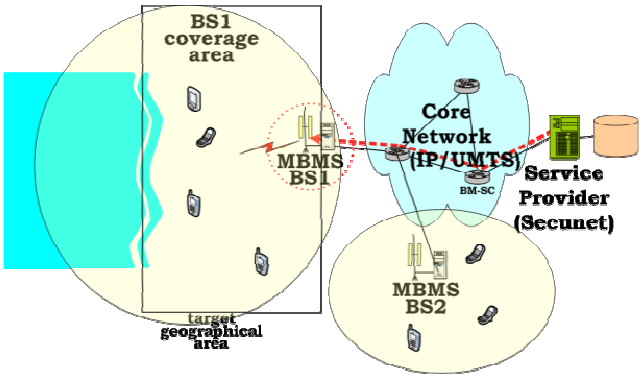


Figure 9. RATCOM Application Scenario with MBMS

C. *Software Defined Networks*

In many contexts, Public Safety Networks (PSN) rely on multi-hop transmissions to deliver the information. This makes the discussion about wireless capacity and efficient use of the spectrum as a central topic for these networks. Some works, such as Tenoc [10], try to reduce the volume of data transmitted into a wireless network using software defined networks or network coding. Network coding [11] is a technique that permits a sensible reduction into the number of data transmissions. We believe that this kind of technique will have a huge impact on the data forwarding in the future. Network coding is a packet dissemination strategy that aims to improve the throughput and increase the robustness of wireless networks. Network coding implements a store, code and forward paradigm, where each node stores the incoming packets in a temporary buffer and at each transmission time, the node sends a combination of the stored data. To successfully decode N packets a node has to collect N independent combinations of packets. Reducing the number of packet transmission to deliver data to multiple destinations is an effective strategy to increase the network throughput.

V. FEASABILITY EXPERIMENTATION

Administrative officials were invited to a system feasibility demonstration at the end of the project [12]. The experimentation featured the downstream component built according to the planned administrative hierarchy, as described below. It concluded the validation phase of the project, which targeted the feasibility and successful operation of the integrated system only, leaving performance analysis for future work. In the setup shown, the Control Centre is directly interconnected with a professional alerting system (for firemen or rescue teams, for example), the SECUNET network, which also hosts a relay allowing to access the various technologies listed in the previous sections of this paper, using either commercial or experimental equipment. The layout of the demonstrator is pictured in Figure 10.

The initial alert is encoded as an XML (Extensible Markup Language) message, whose template is stored in the Alerting Gateway. This server contains a network manager which is made of plug-ins that allows transferring the

message to the various technologies through mail, file transfer or web service, whichever is the adequate format for the technology. When the alert is triggered by the Control Centre, it is first re-formatted by each plug-in and then forwarded to all the relevant networks.

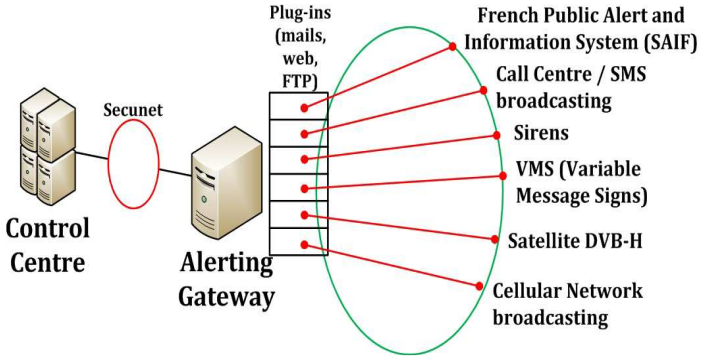


Figure 10. Demonstration of the public Alerting System

During the demonstration, a simulated alert was launched and led within a few seconds to the simultaneous ringing of *(i)* phones registered at the call centre, *(ii)* satellite phones or *(iii)* geographically-scoped mobile phones receiving SMS (Short Message Service). In addition to these existing technologies, the demonstrator was connected to an experimental test setup, showing the operation of multicast over LTE network and virtual Road Side Units, which were described in Section III. The layout of this setup is pictured in Figure 11 and described in more details in [13]. Figure 12 shows that it consisted mainly in laptop computers, running the LTE OpenAirInterface [14] software platform under Linux. When the alert was triggered at the Control Centre, an email was sent to the Alerting Application at the Cellular Network Gateway, appearing on the bottom right of Figure 10 or as "backhaul" in Figure 11. It resulted in the appearance of a pop-up window in each of the end user terminals. This demonstration concluded that existing state of the art and future technologies are capable to be combined in an integrated alerting system, enhancing its effect and efficiency.

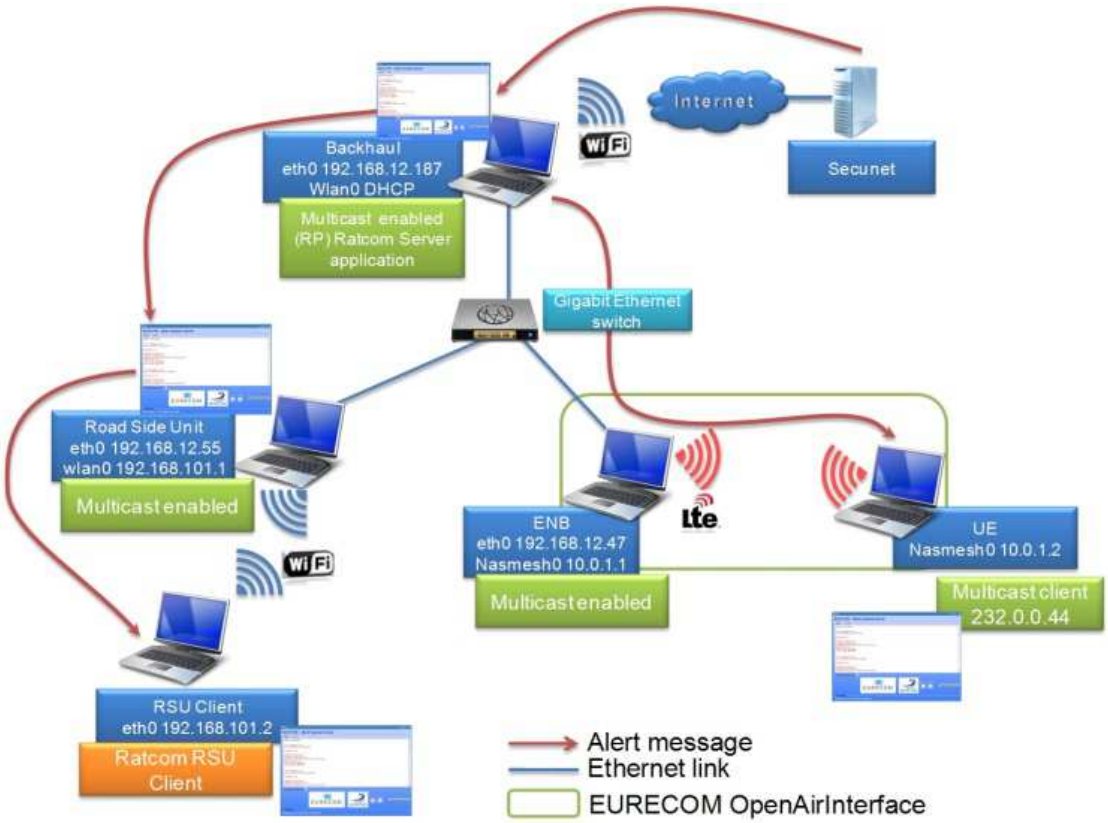


Figure 11. Experimental testbed setup for the demonstration

Figure 12. Picture of the experimental setup during the demonstration

VI. CONCLUSION

In this paper were presented several technologies to be deployed in the medium to long term. Some are completely new (WiMAX, DVB-SH or vehicular networks), others are the future evolution of existing communication networks (Ku band satellite networks, cellular mobile telephony).

Each of these technologies offers specific characteristics and particular interest for the broadcast of alert messages. The DVB-SH satellite broadcast network reaches a large number of users by stopping the radio and TV programs received on fixed or mobile devices, and replace them by the alert bulletin. It does not require the availability of a terrestrial network and therefore does not run the risk of being damaged by a natural disaster. The Ku band satellite network connections can also serve as redundancy to the existing network nodes in the case of failure due to a major problem, enabling the safe operation of critical network nodes or connecting a subnet that was isolated. The WiMAX technology, which is in its early deployment, is based on features close to the Internet. It offers the ability to support a Broadcast / Multicast mode and thus to provide broadcast services in geographical areas that cannot be easily connected with a legacy wired network. The CBS is based on existing cellular networks and deployable at medium term. In countries like Japan, CBS is used for mass message broadcast, even if its setting is somehow problematic. It will be advantageously replaced by the LTE that will achieve permanent connections towards all the terminals with a 4G subscription. Vehicular Networks will allow the broadcast of information from car to car in a specific geographical area. The advantage of this technology lies in the fact that it requires no infrastructure and can reach people while they are travelling. The future evolution of cellular networks is still being defined. With CBS and MBMS technologies, it is possible to broadcast a single message to many users, so at a lower cost from the point of view of radio resources, while capitalizing on a network and a massive penetration rate. The PWS systems take advantage of these features to provide a comprehensive model of early warning network. A proposal to extend the geographical feature of MBMS and increase its efficiency has also been introduced. Finally, the successful demonstration held at the end of the project and showing the downstream public alerting system has been described.

All these technologies can reach in a very limited time a significant number of users and are particularly relevant to a potential connection to the downstream component of a future public warning system, enhancing its overall efficiency. The availability of these technologies in the near or longer future depends mainly on their commercial success, according to business models and the return on investment expected from their deployment. Nevertheless, it is the administrative authorities who ultimately may decide on the development and promote the implementation of the functionalities needed to connect them to a global safety system.

REFERENCES

[1] Offcial RATCOM project web site : http://ratcom.org, last accessed on 30/06/2011 (closed as of 15/12/2012, refer to http://www.afpcn.org/100601VigilanceAlerte/AFPCN-100601-1425AriasBuffardCedralis.pdf)

[2] Michelle Wetterwald, Christian Bonnet, Daniel Camara, Sebastien Grazzini, Jérome Fenwick, Xavier Ladjointe, and Jean-Louis Fondere, "Future Architectures for Public Warning Systems", ICNS 2011, pp. 104-109, 7th International Conference on Networking and Services, May 22-27, 2011, Venice/Mestre, Italy

[3] ETSI TS 102 585 V1.1.2 : "System Specifications for Satellite services to Handheld devices (SH) below 3 GHz"

[4] ETSI TS 102 636-3: "Intelligent Transportation System (ITS); Vehicular Communications; GeoNetworking; Part 3: Network architecture".

[5] Stefano Annese, Claudio Casetti, Carla-Fabiana Chiasserini, Nazario Di Maio, Andrea Ghittino, and Massimo Reineri, “Seamless Connectivity and Routing in Vehicular Networks with Infrastructure”, IEEE Journal on Selected Areas in Communications, Vol. 29, No. 3, pp. 501-514, March 2011

[6] D. Camara, C. Bonnet and F. Filali, “Propagation of Public Safety Warning Messages”, IEEE WCNC 2010, pp. 1-6, Sydney, Australia

[7] 3GPP TS 23.246, “MBMS; ARCHITECTURE AND FUNCTIONAL DESCRIPTION”, V8.3.0 (03-2009)

[8] 3GPP TR 22.268; “Public Warning System (PWS) Requirements”; V9.2.1 (06-2009)

[9] M. Wetterwald, “A case for using MBMS in geographical networking”, ITST 2009, pp 309-313, October 2009, Lille, France.

[10] Stéphane Rousseau, Farid Benbadis, and Damien Lavaux, "Tendoc: A Network Coding Video Transmission for Public Safety," mass, pp.953-954, 2011 IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems, 2011

[11] Rudolf Ahlswede, Ning Cai, Shuo Yen Robert Li, and Raymond W. Yeung, "Network information flow". IEEE Transactions on Information Theory, vol. 46, pp1204-1216, 2000

[12] http://www.lepetitnicois.fr/article/alerte-tsunami-en-m%C3%A9diterran%C3%A9e-%C3%A7a-marche-45857.html

[13] Daniel Câmara, Christian Bonnet, Michelle Wetterwald, and Navid Nikaein, "Multicast and virtual road side units for multi technology alert messages dissemination", WMAPS 2011, 1st International Workshop on Mobile Ad-Hoc Networks for Public Safety Systems, October 21, 2011, Valencia, Spain

[14] http://www.openairinterface.org/

Capacity Evaluation of a New Scheduler with Call Admission Control to Fixed WiMAX Networks with Delay Bound Guarantee

Eden Ricardo Dosciatti
GETIC-NATEC-UTFPR
Federal University of Technology
Pato Branco - Parana - Brazil
Email: edenrd@utfpr.edu.br

Walter Godoy Junior
NATEC-CPGEI-UTFPR
Federal University of Technology
Curitiba - Parana - Brazil
Email: godoy@utfpr.edu.br

Augusto Foronda
DAELN-NATEC-UTFPR
Federal University of Technology
Curitiba - Parana - Brazil
Email: foronda@utfpr.edu.br

***Abstract*—IEEE 802.16, also known as WiMAX, is a solution for mobile and fixed access to broadband networks, currently in development by the Working Group of the Institute of Electrical and Electronics Engineers - IEEE. The WiMAX Working Group focuses on the development of a standard for wireless broadband metropolitan area networks, whose main goal is to allow high-speed access to data, video and voice services. As a wireless broadband technology, WiMAX networks implement Quality of Service (QoS) mechanisms as a crucial element to satisfy users' demands for high data rates. QoS mechanisms and bandwidth allocation are covered by IEEE 802.16 standard. However, the exact details of scheduling and call admission control management, which guarantee QoS as required by multimedia applications, are left unspecified by the standard. In fact, the standard supports scheduling only for fixed-size real-time service flows. The choice of a scheduling algorithm for WiMAX systems is of major importance. A efficient, robust and fair WiMAX scheduling algorithm is still an open issue. Based on these facts, a new scheduler with call admission control with delay bound guarantee was proposed. The new scheduler calculates an optimal time frame, which allows the number of stations allocated in the system to be maximized and manages the delays required by each user. Properties of this algorithm are investigated both theoretically and through simulations. The results show that an upper bound on the delay can be achieved for a large range of network loads, with bandwidth optimization.**

Keywords-IEEE 802.16; WiMAX; QoS; Latency-Rate; scheduling; time frame; call admission control.

I. INTRODUCTION

The deployment of high-speed Internet access is often cited as an open challenge for the second decade of this century. Also known as broadband Internet, it is effective in reducing physical barriers to the transmission of information, as well as transaction costs, and is fundamental in fostering competitiveness. However, providing wired access to broadband Internet is costly and sometimes infeasible, since the investment needed to deploy cabling throughout a region often outweighs the service provider's financial gains. One of the possible solutions in reducing the costs of deploying broadband access in areas where such infrastructure is not present is to use wireless technologies, which require no cabling and reduce both implementation time and cost [2].

This was one of the motivations behind the development by the IEEE (Institute of Electrical and Electronics Engineers) of the 802.16 standard for wireless access [3], also known as Worldwide Interoperability for Microwave Access (WiMAX). It is an emerging technology for next-generation wireless networks, which provides supports for a large number of both mobile and nomadic (fixed) users distributed over a wide geographic area. Furthermore, this technology provides strict QoS (Quality of Service) guarantees for data, voice and video applications [4].

As a service provider, WiMAX creates new alternatives for applications such as telephony, TV broadcasts, broadband Internet access for residential users, and commercial, industrial and university centers. The development of this new market niche represents a revolution for telecommunications companies and interconnection equipment manufacturers [5]. Moreover, WiMAX enables broadband connection for areas, which are inaccessible or lacking in infrastructure, since it requires no complex physical installations of cable connections and traditional technologies [6].

Motivated by the growing need for ubiquitous, high-speed network access, wireless technology is an option to provide a cost-effective solution that may be deployed quickly and easily, providing high bandwidth connectivity in the last mile. However, despite its many advantages, such as low deployment and maintenance costs, ease of configuration, and device mobility, there are challenges that must be overcome in order to further advance its widespread use. The increasing deployment of wireless infrastructure enables a variety of new applications that require flexible, but also robust, support by the network, such as multimedia applications including video streaming and VoIP (Voice over Internet Protocol), which demand real-time data delivery [7].

To this purpose, the IEEE 802.16 standard introduces a set of mechanisms, such as service classes and several coding and modulation schemes that adapt themselves according to channel conditions. However, the standard leaves certain

issues pertaining to network resource management and scheduling algorithms open.

This paper presents a new scheduler with admission control of connections to a WiMAX Base Station (BS). We develop an analytical model based on Latency-Rate (LR) server theory [8], which an ideal frame size, called the Time Frame (TF), is estimated, with guaranteed delays for each user. At the same time, the number of stations allocated in the system is maximized. In this procedure, framing overhead generated by the MAC (Medium Access Control) and PHY (Physical) layers was taken into account when calculating the length of each time slot. After developing this model, a set of simulations is presented for constant bit rate (CBR) and variable bit rate (VBR) streams, with performance comparisons between situations with different delays and different TFs. The results show that an upper limit on the delay may be achieved for a wide range of network loads, thus optimizing bandwidth.

The paper is an extension to [1] and is structured as follows. In Section II, related research is described. In Section III, a brief description of the IEEE 802.16 standard is presented. Our analytical model for packet scheduling is proposed and explained in Section IV. Evaluation of the capacity of the new scheduler with Call Admission Control (CAC) is shown in Section V. Conclusions are presented in Section VI.

II. Related Work

Several scheduling algorithms and QoS architectures for Broadband Wireless Access (BWA) have been proposed in the literature [9-15], since the standard only specifies signaling mechanisms and no specific scheduling and admission control algorithms. However, many of these solutions only address the implementation or addition of a new QoS architecture to the IEEE 802.16 standard. A scheduling algorithm decides the next packet to be served on the queue and is one of the mechanisms responsible for distributing bandwidth among several streams (by assigning each flow the bandwidth that was required and available). In these proposals [9-15], there are often no analytical models for ensuring maximum delay and maximizing the number of SSs (Subscriber Stations) allocated in the system, which are represented accurately by certain performance metrics, such as the delay, of the medium access protocol.

In [9], a packet scheduler for IEEE 802.16 uplink channels based on a hierarchical queue structure is proposed. A simulation model is developed to evaluate the performance of the proposed scheduler. However, despite presenting simulation results, the authors overlooked the fact that the complexity of implementing this solution is not hierarchical, and do not define clearly how requests for bandwidth are made.

In [10], the authors propose a QoS architecture to be built into the IEEE 802.16 MAC sublayer, which significantly impacts system performance, but do not present an algorithm that makes efficient use of bandwidth.

In [11], the authors present a simulation study of the IEEE 802.16 MAC protocol operating with an OFDM (Orthogonal Frequency Division Multiplexing) air interface and full-duplex stations. System performance is evaluated under different traffic scenarios, by varying the values of a set of relevant system parameters. Regarding data traffic, it was observed that the overhead due to the physical transmission of preambles increases with the number of stations.

In [12], a polling-based MAC protocol is presented along with an analytical model to evaluate its performance, considering a system where the BS issues probes in every frame to determine bandwidth requirements for each node. The authors developed closed-form analytical expressions for cases in which stations are polled at the beginning or at the end of uplink subframes. It is not possible to know how the model may be developed to provide delay guarantees.

In [13], the proposal is of a QoS architecture in which the scheduler is based on packet lifetime for each type of flow. The process of data communication between BS and SS is considered from the start, that is, connection and negotiation of traffic parameters such as bandwidth and delay. The proposal features an architecture defined in well-structured blocks, which may make data flows and architecture actions inaccurate. However, despite presenting simulation results, the work neglects performance by not adequately addressing the functional blocks of the proposed architecture and by not specifying clearly how lifetime is calculated for each packet.

In [14], the scheduling algorithm handles traffic with Best Effort (BE), and it is concluded that there exists considerable difficulty in estimating the amount of bandwidth required due to dynamic changes in traffic transmission rate. The purpose of this algorithm is to ensure fairness in bandwidth allocation among BE flows and full bandwidth usage. The system measures the transmission rate for each flow and allocates bandwidth based on the average transmission rate.

Finally, in [15] the author presents a well-established architecture for QoS in the IEEE 802.16 MAC layer. The subject of this work is the component responsible for allocating uplink bandwidth to each SS, although the decision is taken based on the following aspects: the bandwidth required by each SS for uplink data transmission, periodic bandwidth needs for UGS flows in SSs and the bandwidth required for making requests for additional bandwidth.

Considering the limitations exposed above, these works form the basis of a generic architecture, which can be extended and specialized. However, in these studies, the focus is in achieving QoS guarantees, with no concerns for maximizing the number of allocated users in the network. This paper presents a scheduler with admission control of connections to the WiMAX BS. We developed an

analytical model based on Latency-Rate (LR) server theory [8], which an ideal frame size called Time Frame (TF) was estimated, with guaranteed delays for each user and maximization of the number of allocated stations in the system. A set of simulations is presented with constant bit rate (CBR) and variable bit rate (VBR) streams and performance comparisons are made for different delays and different TFs. The results show that an upper bound on the delay may be achieved for a large range of network loads with bandwidth optimization.

A. Latency-Rate Servers

Providing quality of service (QoS) guarantees in a packet network requires the use of traffic scheduling algorithms in the routers. The function of a scheduling algorithm is to select, for each outgoing link of the router, the packet to be transmitted next cycle from the available packets belonging to the sessions sharing the output link.

Since networks are unlikely to be homogeneous in the type of scheduling algorithms employed by the individual routers, a general model for the analysis of scheduling algorithms will be a valuable tool in the design and analysis of such networks.

In work [8] was developed a model to study the behavior of the worst-case of individual sessions in a network of schedulers where the schedulers may employ a broad range of scheduling algorithms. This approach allows to calculate tight bounds on the end-to-end delay of individual sessions and the buffer sizes needed to support them in an arbitrary network of schedulers. The basic approach consists in defining a general class of schedulers, called *Latency-Rate* servers [8], or simply *LR* servers. The theory of *LR* servers provides a means to describe the worst-case behavior of a broad range of scheduling algorithms in a simple and elegant manner. This theory is based on the concept of a busy period of a session, a period of time during which the average arrival rate of the session remains at or above its reserved rate r_i. For a scheduling algorithm to belong to the *LR* class, it is only required that the average rate of service offered by the scheduler to a busy session, over every interval starting at time θ from beginning of the busy period, is at least equal to its reserved rate. The parameter θ is called latency of the scheduler.

The behavior of an LR scheduler is determined by two parameters: the *latency* (θ) and the *allocated rate* (r_i). The latency of LR server may be seen as the worst-case delay seen by the first packet of the busy period of a session, which is a packet arriving when the queue is empty session. The latency of a particular scheduling algorithm may depend on its internal parameters, its transmission rate on the outgoing link, and the allocated rates of various sessions. However, the maximum end-to-end delay experienced by a packet in a network of schedulers can be calculated from only the latencies of the individual schedulers on the path of the session, and the traffic parameters of the session that generated the packet. Since the maximum delay in a scheduler increases directly in proportion to its latency, the model brings out the significance of using low-latency schedulers to achieve low end-to-end delays. Likewise, upper bounds on the queue size and burstiness of individual sessions at any point within the network can be obtained directly from the latencies of the schedulers.

III. The IEEE 802.16 standard

The basic topology of a IEEE 802.16 network includes two entities that participate in the wireless link: Base Stations (BS) and Subscriber Stations (SS), as shown in Figure 1 [16].

The BS is the central node, responsible for coordinating communication and providing connectivity to SSs. BSs are kept in towers distributed so as to optimize network coverage area, and are connected to each other by a backhaul network, which allows SSs to access external networks or exchange information between themselves.

Networks based on the IEEE 802.16 standard can be structured in two schemes. In PMP (Point-to-multipoint) networks, all communication between SSs and other SSs or external networks takes place through a central BS node. Thus, traffic flows only between SSs and the BS (see Figure 1). In Mesh mode, SSs communicate with each other without the need for intermediary nodes; that is, traffic can be routed directly through SSs. Thus, all stations are peers, which can act as routers and forward packets to neighboring nodes [17]. This article only considers the PMP topology, since it is implemented by first-generation WiMAX devices,

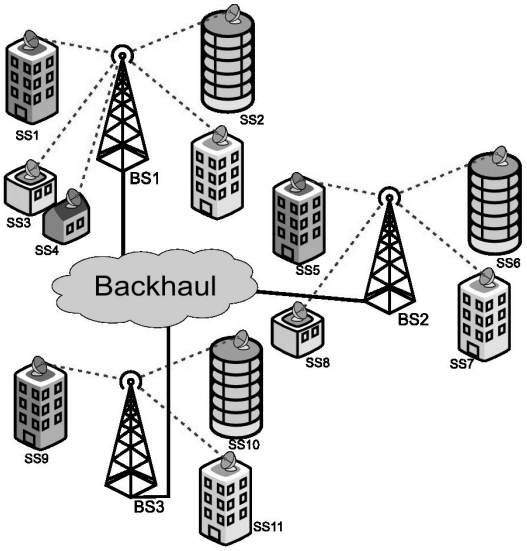


Figure 1. IEEE 802.16 Network Architecture

and also due to the strong trend towards its adoption by Internet providers because it allows them to control network parameters in a centralized manner, without the need to recall all subscriber stations [5].

Although it is referred to as fixed pattern, IEEE 802.16 allows stations to provide customers with low-speed mobility. A feature missing in this pattern and that justifies its designation as fixed is the possibility of performing handoffs/handovers, which allow a client station to switch to another base station without losing connectivity. In this case, subscriber stations are instead called mobile stations. The functionality of handoff/handover was included in the IEEE 802.16 standard in early 2006 with the publication of the IEEE 802.16e [18], which quickly received the name of "IEEE 802.16 mobile".

WiMAX technology can reach a theoretical maximum distance of 50 km [19]. Data transmission rates can vary from 50 to 150 Mbps, depending on channel frequency band width and modulation type [20]. Communication between a BS and SSs occurs in two different channels: uplink (UL) channel, which is directed from SSs to the BS, and downlink (DL) channel, which is directed from the BS to SSs. DL data is transmitted by broadcasting, while in UL access to the medium is multiplexed. UL and DL transmissions can be operated in different frequencies using Frequency Division Duplexing (FDD) mode or at different times using Time Division Duplexing (TDD) mode.

In TDD, the channel is segmented in fixed-size time slots. Each frame is divided into two subframes: a DL subframe and an UL subframe. The duration of each subframe is dynamically controlled by the BS; that is, although a frame has a fixed size, the fraction of it assigned to DL and UL is variable, which means that the bandwidth allocated for each of them is adaptive. Each subframe consists of a number of time slots, and thus both the SSs and the BS must be synchronized and transmit the data at predetermined intervals. The division of TDD frames between DL and UL is a system feature controlled by the MAC layer. Figure 2 [10] shows the structure of a TDD frame. In this paper, the system was operated in TDD mode with the OFDM (Orthogonal Frequency Division Multiplexing) air interface, as determined by the standard [3].

IV. Analysis of the Analytical Model

A minimum acceptable performance level should be sought throughout the development of any system, be it computer-related or not. This requires a measure or gauge of performance in these systems. To accomplish this, there exist design tools that provide the analyst with different metrics and measures. Within this scope, some related system characteristics are proposed and discussed in this article. To accomplish this, this section presents an analytical model of the new scheduler and an analytical description of its call admission control facility.

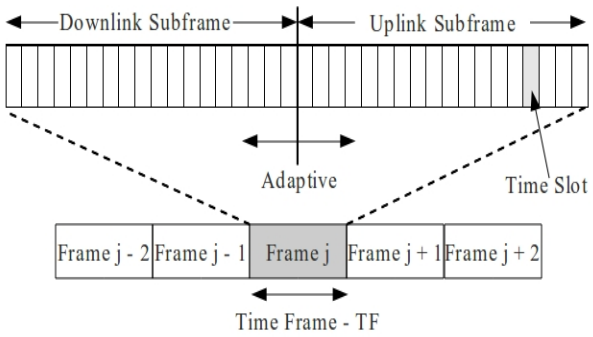


Figure 2. IEEE 802.16 Frame Structure

A. System Description

Figure 3 [21] illustrates a wireless network operating the newly proposed scheduler with connection admission control, which is based on a modified LR scheduler [8] and uses the token bucket algorithm.

The basic approach consists on the token bucket limiting input traffic and the LR scheduler providing rate allocation for each user. Then, if the rate allocated by the LR scheduler is larger than the token bucket rate, the maximum delay may be calculated.

A scheduler that provides guaranteed bandwidth can be modeled as an LR scheduler. The behavior of an LR scheduler is determined by two parameters for each session i: latency θ_i and allocated rate r_i. The latency θ_i of the scheduler may be seen as the worst-case delay and depends on network resource allocation parameters. In the new scheduler with call admission control, the latency θ_i is a TF period, which is the time needed to transmit a

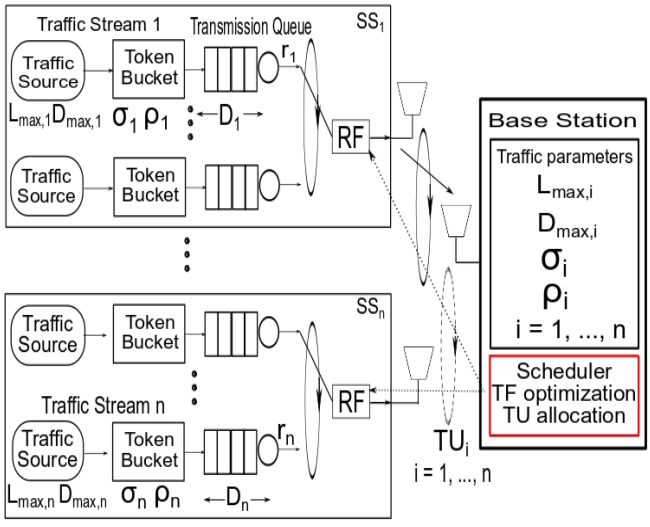


Figure 3. Wireless Network with New Scheduler

maximum-size packet and separation gaps (TTG and RTG) of DL and UL subframes. In the new scheduler, considering the delay for transmitting the first packet, the latency θ_i is given by

$$\theta_i = T_{TTG} + T_{RTG} + T_{DL} + T_{UL} + \frac{L_{max,i}}{R} \quad (1)$$

where T_{TTG} and T_{RTG} are the DL and UL subframe gap durations, T_{DL} and T_{UL} are the DL and UL subframe durations, $L_{max,i}$ is the maximum packet size and R is the outgoing link capacity.

Now, we show how the allocated rate r_i for each session i may be determined, and how to optimize TF in order to increase the number of connections accommodated.

B. CAC Description

An LR scheduler can provide a bounded delay if input traffic is shaped by a token bucket. A token bucket [2] is a non-negative counter, which accumulates tokens at a constant rate ρ_i until the counter reaches its capacity σ_i. The rate of incoming packets (ρ_i) is constant because the parameters of the token bucket, for all three types of traffic, that will be used for performance evaluation are constant, i.e., audio will be 64 kb/s, VBR video will be 500 kb/s, and MPEG4 video will be 4100 kb/s. Packets from session i can be released into the queue only after removing the required number of tokens from the token bucket. In an LR scheduler, if the token bucket is empty, arriving packets are dropped; however, our model ensures that there will always be tokens in the bucket and that no packets are dropped, as described in Section IV. If the token bucket is full, a maximum burst of σ_i packets can be sent to the queue. When the flow is idle or running at a lower rate as the token size reaches the upper bound σ_i, accumulation of tokens will be suspended until the arrival of the next packet. We assume that the session starts out with a full bucket of tokens. In our model, we consider IEEE 802.16 standard overhead for each packet. Then, as we will show below, the token bucket size will decrease by both packet size and overhead.

The application using session i declares the maximum packet size $L_{max,i}$ and requires maximum allowable delay $D_{max,i}$, which are used by the WiMAX scheduler to calculate the service rate for each session so as to guarantee the required delay and optimize the number of stations in the network. Incoming traffic $A_i(t)$ from session $i (i = 1, ..., N)$ passes through a token bucket inside the user terminal during the time interval $(0, t)$.

This passage of data traffic by the token bucket is bounded by

$$A_i(t) \leq \sigma_i + \rho_i t \quad (2)$$

where σ_i is the bucket size and ρ_i is the bucket rate.

Then, the packet is queued in the station until it is transmitted via the wireless medium. Queue delay is measured as the time interval between the receipt of the last bit of a packet and its transmission. In the new scheduler with call admission control, queuing delay depends on token bucket parameters, network latency and allocated rate. In [8] and [22], it is shown that if input traffic $A_i(t)$ is shaped by a token bucket and the scheduler allocates a service rate r_i, then an LR scheduler can provide a bounded maximum delay D_i:

$$D_i \leq \frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i} \quad (3)$$

where σ_i is the token bucket size, r_i is the service rate, θ_i is the scheduler latency, $\frac{L_{max,i}}{r_i}$ is the maximum size of a package and, $\frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i}$ is the bound on the delay, D_{bound}.

Equation (3) is an improved bound on the delay for LR schedulers. Thus, the token bucket rate plus the overhead transmission rate must be smaller than the service rate to provide a bound on the delay. The upper bound D_{bound} should be smaller than or equal to the maximum allowable delay:

$$\frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i} \leq D_{max,i} \quad (4)$$

Therefore, three different delays are defined. The first is the maximum delay D_i, the second is the upper bound on the delay D_{bound} and the third is the required maximum allowable delay $D_{max,i}$. The relation between them is $D_i \leq D_{bound} \leq D_{max,i}$.

So, the delay constraint condition of the new scheduler is

$$\frac{(\sigma_i' - L_{max,i}')TF}{r_i'TF - \Delta R + L_{max,i}'} + TF + \\ + \frac{L_{max,i}'}{R} + T_{TTG} + T_{RTG} \leq D_{max,i} \quad (5)$$

where σ_i' is the token bucket size with overhead, $L_{max,i}'$ is the maximum size of a packet with overhead (preamble+pad), TF is the time frame, r_i' is the rate allocated by the server with overhead, R is the outgoing link capacity, T_{TTG} is the gap between downlink and uplink subframes, T_{RTG} is the gap between uplink and downlink subframes, $D_{max,i}$ is the maximum allowable delay and Δ is the sum of initial ranging and BW request, which is the uplink subframe overhead and whose value will be discussed when evaluating performance. Physical rate, maximum packet size and token bucket size are parameters declared by the application. However, TF and total allocated service rate must satisfy Equation (5).

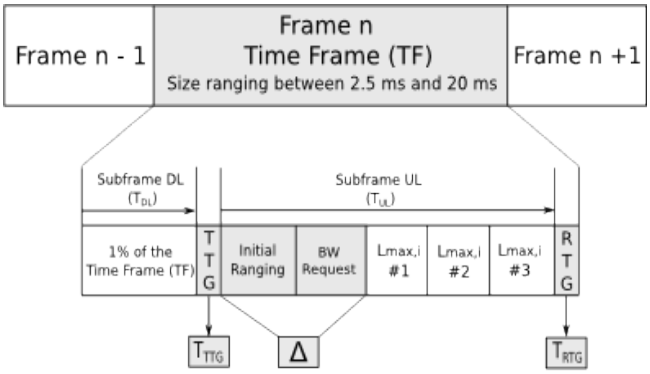


Figure 4. Frame structure with TDD allocation formulas of Equation (5)

Figure 4 shows a frame structure with TDD allocation formulas as described by Equation (5).

The second delay constraint condition to TF and service rate is that the token bucket rate plus the rate to transmit overhead and a maximum-sized packet must be smaller than the service rate to place a bound on delay. Thus, the second constraint condition is

$$\rho_i + \frac{\Delta R + L'_{max,i}}{TF} \leq r'_i \quad (6)$$

where ρ_i is the bucket rate, Δ is the uplink subframe overhead, R is the outgoing link capacity, $L'_{max,i}$ is the maximum packet size with overhead, TF is the time frame and r'_i is the rate allocated by the service with overhead.

Previous schedulers do not provide any mechanisms to estimate the TF needed to place a bound on delay or to maximize the number of stations, because each application requires a TF without the use of criteria to calculate the time assigned to each user. However, TF estimation is important because of a tradeoff. A small TF reduces maximum delay, but increases overhead at the same time. On the other hand, a large TF decreases overhead, but increases delay. Therefore, we must calculate the optimal TF to allocate the maximum number of users under both constraints. The maximum number of users is achieved when the service rate for each user is the minimum needed to guarantee the bound on the delay, D_{bound}.

To find the maximum number of users in each frame, we solve a problem of non-linear optimization. Solving non-linear problems is characterized by not having a single algorithm for solving their problems. The biggest difficulty with this approach is the uncertainty that the solution to this problem is really the best, and this is a fact inherent in the non-linear nature of the problem, whereas its great advantage is the scope, that is, once the mathematical model developed the problem, with its objective function and its constraints, usually no simplification is needed in terms of formulation.

So, in this work, nonlinear optimization makes use of search techniques using numerical information given in an iterative process, generating better solutions in the optimization process. These techniques allow us to use numerical methods to solve problems when there is no known analytical solution.

In the specific case of this work, an approach step-by-step was used, where a small initial value for the TF is determined, in this case, the value of 2.5 ms (lower reference value for the TF according to 802.16 standard [3]). After, the value of r'_i is calculated and the process is repeated with a certain step length, in this case, 0.5 ms, until the minimum value of r'_i is found, satisfying the constraints of Equations (5) and (6). The value of the step length can be determined randomly by the limit of 20 ms (maximum value of a frame in accordance with 802.16 standard [3]) and there will always be a solution because at every step the two constraints of Equations (5) and (6) will be confronted in order to verify that the minimum value r'_i found.

V. Performance Analysis

To analyze the IEEE 802.16 MAC protocol behavior with respect to the new scheduler with call admission control, this section presents numerical results obtained with the analytical model proposed in the previous section. Then, with a simulation tool, the proposed analytical model is validated by showing that the bound on the maximum delay is guaranteed. In this section, two types of delays are treated: required delay, in which the user requires the maximum delay, and the guaranteed maximum delay, which is calculated with the analytical model.

A. Calculation of Optimal Time Frame

In this paper, the duration of downlink subframes is fixed at 1% of the TF because our interest is only in the uplink subframe. In the simulation, after finding the optimal number of SSs per frame for each traffic flow, the header value of the uplink subframe is calculated at a rate of 10% of the value of an OFDM symbol [2].

All PHY and MAC layer parameters used in simulation are summarized in Table I.

Performance of the new scheduler with call admission

Table I
PHY and MAC parameters

Parameter	Value
Bandwidth	20 MHz
OFDM Symbol Duration	13,89 μs
Delay	5, 10, 15 and 20 ms
Δ (Initial Ranging and BW Request) = 9 OFDM Symbols	125,10 μs
TTG + RTG = 1 OFDM Symbol	13,89 μs
UL Subframe (preamble + pad) = 10% OFDM Symbol	1,39 μs
Physical Rate	70 Mbps
DL Subframe	1% TF

Table II
Token bucket parameters

	Audio	**VBR video**	**MPEG4 video**
Token Size (bits)	3000	18000	10000
Token Rate (kb/s)	64	500	4100

control is evaluated as the delay requested by the user and assigned stations. Station allocation results, in the system with an optimal TF, limited by the delay requested by the user, are described in sequence. The first step is defining token bucket parameters, which are estimated according to the characteristics of incoming traffic and are listed on Table II. It's worth noting that the details about the traffic must be known in advance. This is normal for various applications such as audio, CBR and video on demand.

Thus, the optimal TF value is estimated according to the PHY and MAC layer's parameters (see Table I), token bucket parameters (see Table II), required maximum allowable delay, physical rate and maximum packet size. With all parameters defined, and with the constraints set by Equations (5) and (6), described in Section IV-B, we use a step-by-step approach, starting with a small TF of 2.5 ms, calculating r_i' and repeating this process every 0.5 ms until the minimum r_i' that satisfies both equations is found. The graph in Figure 5 shows the optimal TF value, for four delay values required by users (5, 10, 15 and 20 ms):

- For a requested delay of 5 ms, the optimal TF is 3 ms.
- For a requested delay of 10 ms, the optimal TF is 6.5 ms.
- For a requested delay of 15 ms, the optimal TF is 10.5ms.
- For a requested delay of 20 ms, the optimal TF is 15 ms.

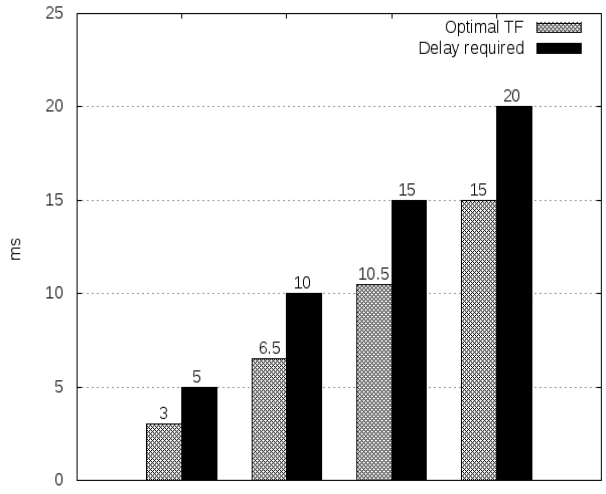


Figure 5. Optimal TF

Figure 6 shows the number of SSs assigned to each traffic type in each frame, through of the optimal TF calculated. The result shows the maximum number of SSs assigned to each range of optimal TF values for each traffic type. It should be noted that three traffic types were used: audio traffic, VBR video traffic and MPEG4 video traffic. For the simulation, the allocation of users is performed by traffic type; i.e., only one traffic at a time will be transmitted within each frame.

As an example, Figure 6d shows that when the user-requested delay is of 20 ms, an optimal TF of 15 ms is calculated and 50 users can be allocated for audio traffic, or 30 users for VBR video traffic, or 13 users for MPEG4 video traffic.

Two important observations from Figure 6d should be highlighted:

1) With a requested delay of 20 ms, we cannot choose a TF of less than 15 ms, since the restrictions placed by Equation (5) (which regards delay) and Equation (6) (which regards the token bucket) are not respected and thus no bandwidth allocation guarantees exist.
2) We also cannot choose a TF greater than 15 ms, even though it complies with Equations (5) and (6) with respect to guaranteed bandwidth, because there will be a decrease in the number of users allocated to each traffic flow due to increasing delay.

Thus, it is evident that since the IEEE 802.16 standard does not specify an ideal time frame (TF) duration, this approach becomes advantageous because, in addition to meeting the restrictions of the analytical model, it optimizes the allocation of users on the system. The same philosophy holds true for other delay values of 5, 10 and 15 ms.

B. Comparison of User Allocation and Optimal Time Frame

In this work, an optimal TF was reached, so that the number of SSs in the network may be optimized and a maximum delay may be guaranteed. To make a comparison of the results in this work, Figure 7 shows that, for an audio traffic and a requested delay of 15 ms, an optimal TF of 10.5 ms is obtained and 41 users can be allocated. When compared to other randomly-chosen TFs, it may be observed that the optimal TF yields a greater number of users. Thus, when an user requests a delay guarantee, an optimal TF is calculated in order to allocate the largest number of users in a given traffic flow, as seen in the example in Figure 7. It may be noticed, then, that choosing a non-optimal TF will lead to a decreased number of allocated SSs. Therefore, the new scheduler with call admission control proposed herein maximizes the number of SSs in place and ensures an upper bound on maximum delay, as discussed below.

C. Guaranteed Maximum Delay

In this article, only UL traffic is considered. To test the new scheduler's performance, we have carried out

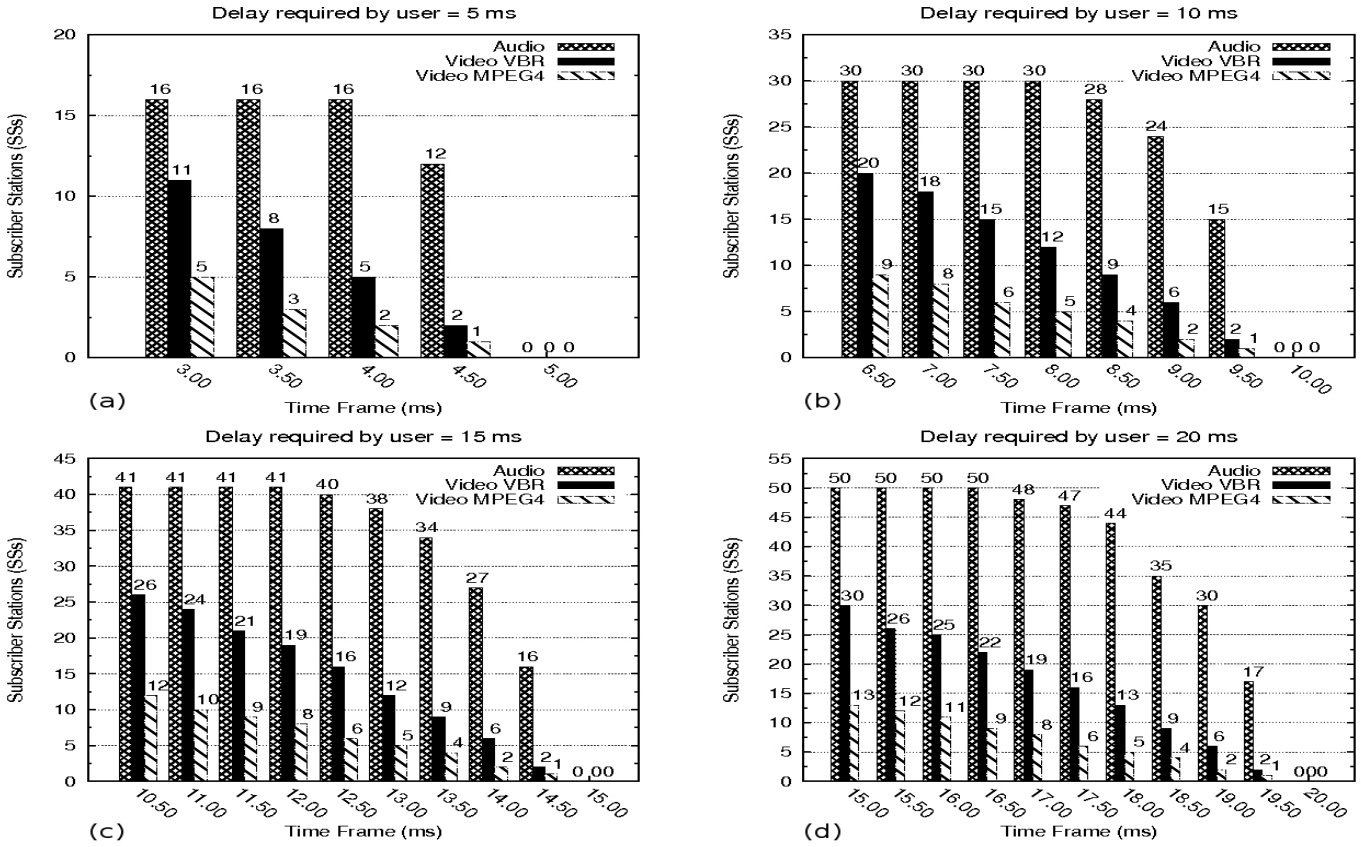


Figure 6. Number of subscriber stations for all delay required

simulations of an IEEE 802.16 network consisting of a BS that communicates with eighteen SSs, with one traffic flow type by SS and the destination of all flows being the BS, as shown in Figure 8. In this topology, six SSs transmit on-off CBR audio traffic (64 kb/s), six transmit CBR MPEG4 video

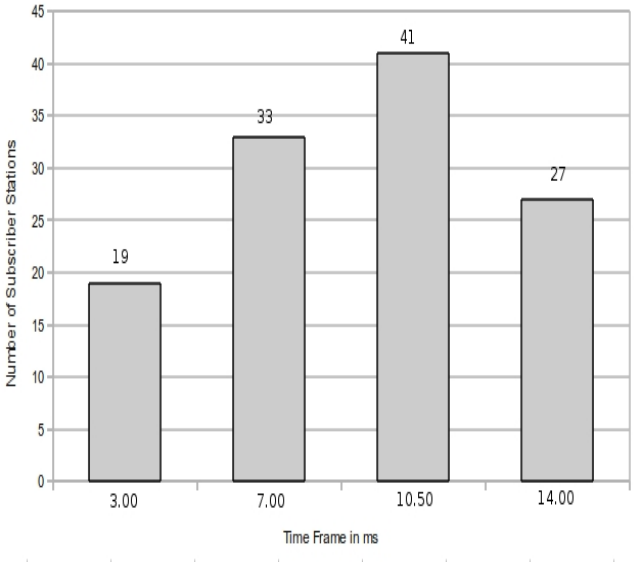


Figure 7. Users assigned as a function of TF, for audio traffic

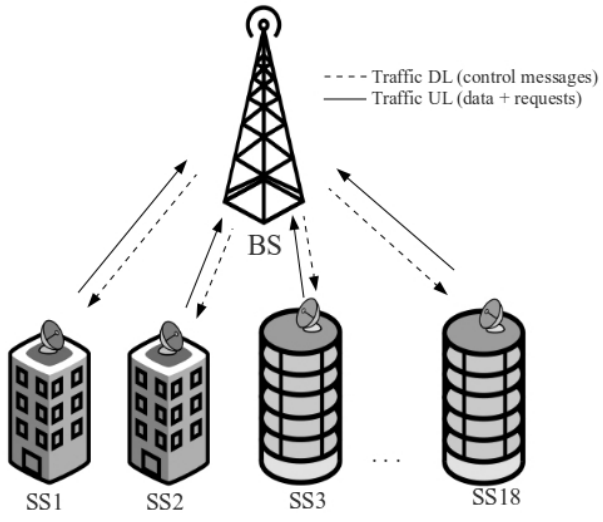


Figure 8. Simulation scenario

Table III
Description of traffic types

Node	**Application**	**Arrival Period** *(ms)*	**Packet Size** *(max) (B)*	**Sending Rate** *(kb/s) (mean)*
$1 \rightarrow 6$	Audio	4.7	160	64
$7 \rightarrow 12$	VBR video	26	1024	$\approx$ 200
$13 \rightarrow 18$	MPEG4 video	2	800	3200

traffic (3.2 Mb/s) and six transmit VBR video traffic. Table III summarizes the different types of traffic used in this simulation.

In Section V-D we have the algorithm of the simulator and its source code, developed in C programming language [23]

In Figure 9, with an optimal TF of 3 ms and an user-requested delay of 5 ms, the average guaranteed maximum delay for audio traffic is 1.50 ms. For VBR video traffic, whose packet rate is variable, the average maximum delay is 1.97 ms. For MPEG4 video traffic, the average maximum delay is 2.00 ms. Data that supports the stated maximum guaranteed delay values is listed in the tables below, which relate the number of packets read in each simulation to the resulting guaranteed maximum delay. A number of simulations were run for each type of traffic to keep results from varying too widely. Our choice of six simulations for each case produced values with noticeably little variation. After running simulations for each optimal TF and each traffic type, averages of resulting guaranteed maximum delays were taken and the graph of Figure 9 was constructed.

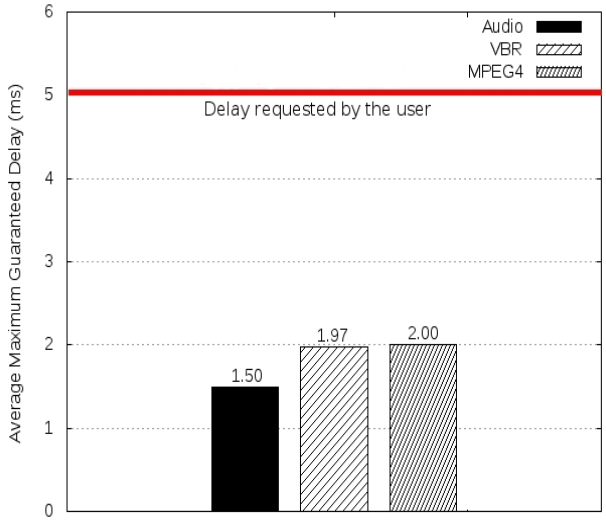


Figure 9. Guaranteed Maximum Delay

Table IV
Algorithm to compute the delay

Step 1	(**initialization**): Initialize the variables of total packet and time frame.
Step 2	(**perform**): **While** the total number of packets is smaller the number of packets in the system **and** the time frame is than smaller than the number of packets in the system, **do**: a. calculate the size of the package. b. calculate the time frame.
Step 3	(**testing**): **If** time frame is greater than the packet size, calculate the packet delay and the total delay, **Else** increment the time frame.
Step 4	(**results**): calculate the average delay and print the result on the screen.

D. Pseudocode of the Algorithm Simulator

In this section, we describe the structure of the simulator and the pseudocode of the simulator. The C programming language [23] was used to build the simulator that calculates the guaranteed maximum delay. In Table IV, is shown the pseudocode algorithm of the simulator to calculate the guaranteed maximum delay. This algorithm uses the parameters of Table III in Section V-C.

After reading the file with the amount of packets, variables that calculate the packet size and time frame, are used to perform the calculation of the delay of each packet, and if the value of packet size calculated is greater than the value of time frame calculated, there was a delay and it will be stored in variables to calculate the delay of each packet and total delay. In the end, the average delay is calculated and print the results on the screen.

This code is generic and is used to calculate the delay of all traffic used in this work.

The tables below show the result of using the simulation algorithm with the three traffic types used, namely audio, VBR video and MPEG4 video.

Table V shows the results of audio traffic simulations. Table VI shows the results of VBR video traffic simulations, whose packet rate is variable. Table VII shows the results of MPEG4 video traffic simulations.

Table V
Audio Traffic

Delay by the user	5 ms	10 ms	15 ms	20 ms
Optimal TF	3 ms	6.5 ms	10.5 ms	15 ms
Packages read *amount*	**Guaranteed Maximum Delay** (ms)			
1000	1.48	3.23	5.24	7.49
3000	1.49	3.24	5.24	7.50
5000	1.49	3.25	5.25	7.50
10000	1.50	3.25	5.25	7.50
30000	1.50	3.25	5.28	7.50
50000	1.50	3.35	5.29	7.51
Mean	1.50	3.25	5.26	7.50
Standard Deviation	0.00816	0.00837	0.02137	0.00632

Table VI
VBR Video Traffic

Delay by the user	5 ms	10 ms	15 ms	20 ms
Optimal TF	3 ms	6.5 ms	10.5 ms	15 ms
Packages read *amount*	**Guaranteed Maximum Delay** (ms)			
2176	2.06	3.48	5.50	7.98
1358	1.94	3.52	5.45	7.96
1177	1.97	3.48	5.59	8.07
1226	2.02	3.32	5.41	8.07
1159	1.87	3.33	5.57	8.08
1449	1.96	3.45	5.53	8.04
Mean	1.97	3.43	5.51	8.03
Standard Deviation	0.06573	0.08438	0.06940	0.05125

Table VII
MPEG4 Video Traffic

Delay by the user	5 ms	10 ms	15 ms	20 ms
Optimal TF	3 ms	6.5 ms	10.5 ms	15 ms
Packages read *amount*	**Maximum Guaranteed Delay** (ms)			
1000	2.00	3.50	5.51	8.01
3000	2.00	3.50	5.50	8.00
5000	2.00	3.50	5.50	8.00
10000	2.00	3.50	5.50	8.00
30000	2.00	3.50	5.50	8.00
50000	2.00	3.50	5.50	8.00
Mean	2.00	3.50	5.50	8.00
Standard Deviation	0.0	0.0	0.00408	0.00408

E. Comparison with other Schedulers

The new scheduler with call admission control, here called *New Scheduler*, was compared to those of [12], here called *Scheduler_1*, and [9], here called *Scheduler_2*. The comparison was accomplished through the ability to allocate users in a particular time frame (TF). Table VIII shows the parameters used for comparisons.

In the graph of Figure 10, we compare the *New Scheduler* with the *Scheduler_1*. A maximum delay of 0.12 ms was requested by the user, and the duration of each frame (TF) was set at 5 ms. Other parameters are listed in Table VIII. In comparison, the *New Scheduler* allocates 28 users in each frame, while the *Scheduler_1*, allocates 20 users. Thus, the *New Scheduler* presents a gain in performance of 40% when compared with the *Scheduler_1*.

In the graph of Figure 11, we compare the *New Scheduler* with the *Scheduler_2*. A maximum delay of 20 ms was

Table VIII
Parameters used for comparisons

Parameter	*Scheduler_1*	*Scheduler_2*
Bandwidth	20 MHz	20 MHz
OFDM symbol duration	13.89 μs	13.89 μs
Delay Requested by the user	0.12 ms	20 ms
Time Frame (TF)	5 ms	10 ms
Maximum Data Rate	70 $Mbps$	70 $Mbps$
Traffic type	Audio	Audio

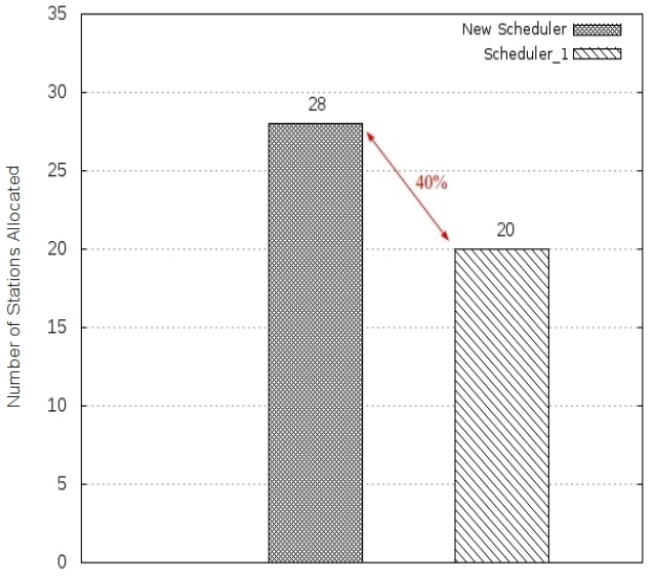


Figure 10. Comparison of user allocation with *Scheduler_1*

requested by the user, and the duration of each frame (TF) was set at 10 ms. Other parameters are listed in Table VIII.

The comparison was extended by also considering frame duration values of 7.00 ms, 8.00 ms and 9.00 ms to demonstrate the efficiency of the *New Scheduler*. For a TF of 10 ms, the *New Scheduler* allocates 41 users in each frame, while the *Scheduler_2* allocates only 33 users. This represents 24.24% better performance for the *New Scheduler*. Similarly, the *New Scheduler* also allocates more users per frame in comparison with the *Scheduler_2* for all other frame duration values.

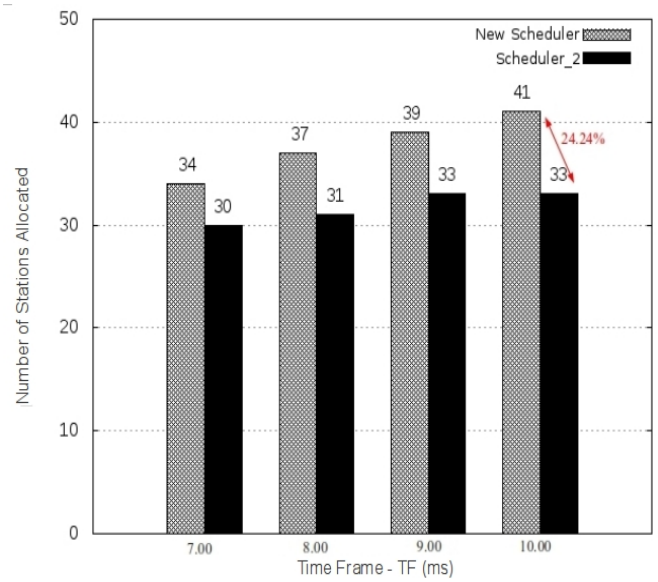


Figure 11. Comparison of user allocation with *Scheduler_2*

VI. Conclusion and Future Work

This work has presented the design and evaluation of a new scheduler with call admission control for IEEE 802.16 broadband access wireless networks (known worldwide as WiMAX) that guarantees different maximum delays for traffic types with different QoS requisites and optimizes bandwidth usage.

A. Conclusion

Firstly, we developed an analytical model to calculate an optimal TF, which allows an optimal number of SSs to be allocated and guarantees the maximum delay required by the user. Then, a simulator was developed to analyze the behavior of the proposed system.

To validate the model, we have presented the main results obtained from the analysis of different scenarios. Simulations were performed to evaluate the performance of this model, demonstrating that an optimal TF was obtained along with a guaranteed maximum delay, according to the delay requested by the user. Thus, the results have shown that the new scheduler with call admission control successfully limits the maximum delay and maximizes the number of SSs in a simulated environment.

B. Future Work

In a communication system with a wireless link, the channel effects can heavily degrade the system performance since the wireless link is time-varying and may experience multipath fading and interference. In future work, the effects of the channel will be treated.

Furthermore, most four improvements will be introduced in order to improve traffic in Fixed WiMAX Networks:

1) The loss of packets in the communication channel will be dealt with so we can get more accurate results.
2) The Call Admission Control (CAC) will use an optimization tool that can perform more efficiently, control of connections that will be served by the system.
3) To calculate the time frame (TF), Particle Swarm Optimization (PSO) [24] is used.
4) The Network Simulator NS-3 [25] is used for the simulations of performance evaluation of these new improvements.

In a communication system with a wireless link, the channel effects can heavily degrade the system performance since the wireless link is time-varying and may experience multipath fading and interference.

Acknowledgment

We thank all researchers and collaborators at the Advanced Nucleus of Communication Technology at UTFPR.

References

[1] E. R. Dosciatti, W. Godoy Jr., and A. Foronda, "New Scheduler with Call Admission Control (CAC) for IEEE 802.16 Fixed with Delay Bound Guarantee," in *Proc. of the First International Conference on Mobile Services, Resources, and Users (MOBILITY 2011)*, Barcelona, Spain, Oct. 2011, pp. 139-145.

[2] A. Gosh, D. Wolter, J. Andrews, and R. Chen, "Broadband wireless access with WiMAX/802.16: current performance benchmarks and future potential," in *IEEE Communications*, v. 43(2), Feb. 2005, pp. 129-136.

[3] IEEE 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems," *IEEE Std., Rev. IEEE Std802.16-2004*, New York, Oct. 2004.

[4] E. G. Camargo, C. B. Both, R. Kunst, L. Z. Granville, and J. Rochol, "Uma Arquitetura de Escalonamento Hierárquica para Transmissões Uplink em Redes WiMAX Baseadas em OFDMA," in *Proc. of Brazilian Symposium on Computer Networks and Distributed Systems (SBRC 2009)*, Recife, Brazil, May 2009, pp. 525-538. (in Portuguese).

[5] C. Eklund, R. B. Marks, K .L. Stanwood, and S. Wang, "IEEE Standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access," in *IEEE Communications Magazine*, v. 40(6), June 2002, pp. 98-107.

[6] WiMAX Forum. *WiMAX Forum*. 2012. [Online]. Available: http://www.wimaxforum.org. [Accessed: Dec. 10, 2012].

[7] Y. Sun, I. Sheriff, E. M. Belding-Royer, and K. C. Almeroth, "Experimental Study of Multimedia Traffic Performance in Mesh Networks," in *Proc. of the International Workshop on Wireless Traffic Measurements and Modeling*, Seattle, EUA, June 2005, pp. 25-30.

[8] D. Stiliadis and A. Varma, "Latency-Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms," in *IEEE-ACM Transactions on Networking*, v. 6(5), Oct. 1998, pp. 611-624.

[9] K. Wongthavarawant and A. Ganz, "Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems," in *International Journal of Communications Systems*, v. 16, Feb. 2003, pp. 81-96.

[10] G. Chu, D. Wang, and S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA System," in *Proc. of IEEE Conference on Communications, Circuits, and Systems*, v. 1, Chengdu, China, June/July 2002, pp. 435-439.

[11] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support," in *IEEE Transactions on Mobile Computing - TMC07*, v. 6(1), Jan. 2007, pp. 26-38.

[12] R. Iyengar, P. Iyer, and B. Sikdar, "Delay Analysis of 802.16 Based Last Mile Wireless Networks," in *Proc. of IEEE Global Telecommunications Conference - GLOBECOM'05*, v. 5, St. Louis, EUA, Dec. 2005, pp. 1-5.

[13] D. Cho, J. Song, M. Kim, and K. Han, "Performance Analysis of the IEEE 802.16 Wireless Metropolitan Area Network," in *IEEE Computer Society, DFMA'05*, Feb. 2005, pp. 130-137.

[14] S. Kim and I. Yeom, "TCP-aware Uplink Scheduling for IEEE 802.16," in *IEEE Communications Letters*, v. 11(2), Feb. 2007, pp. 146-148.

[15] S. Maheshwari, "An Efficient QoS Scheduling Architecture for IEEE 802.16 Wireless MANs," *Master Degree*, K. R. School of Information Technology, Bombay, India, Jan. 2005.

[16] E. R. Dosciatti, W. Godoy Jr., and A. Foronda, "Scheduling Mechanisms with Call Admission Control (CAC) and an Approach with Guaranteed Maximum Delay for Fixed WiMAX Networks," in *Quality of Service and Resource Allocation in WiMAX*, INTECH, Croatia, Feb. 2012, pp. 59-84.

[17] I. F. Akyildiz and X. Wang, "A Survey on Wireless Mesh Networks," in *IEEE Communications Magazine*, v.43(9), Sept. 2005, pp. 523-530.

[18] IEEE 802.16e-2005, "IEEE Standard for Local and Metropolitan Area Networks. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands: Part 16: Air Interface for Fixed Broadband Wireless Access Systems," New York, Dec. 2005.

[19] A. S. Tanenbaum, *Computer Networks*. 4. ed. New Jersey: Prentice-Hall, 2003.

[20] INTEL. *Deploying License-Exempt WiMAX Solutions: White paper*. 16 p., Jan. 2005.

[21] A. Foronda, Y. Higuchi, C. Ohta, M. Yoshimoto, and Y. Okada, "Delay Guarantee and Service Interval Optimization for HCCA in IEEE 802.11e WLANs," in *IEEE Wireless Communications and Networking Conference - WCNC 2007*, v. 1, Hong Kong, Mar. 2007, pp. 2080-2085.

[22] A. Parekh and R. Gallager, " A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: the Single-Node Case," in *IEEE/ACM Transactions Networking*, v. 1(3), June 1993, pp. 344-357

[23] D. M. Ritchie and B. W. Kernighan, *The C Programming Language*. 2. ed. New Jersey: Prentice-Hall, 1988.

[24] J. Kennedy and R. C. Eberhart, "Particle Swarm Optimization," in *Proc. of IEEE International Conference on Neural Networks*, v. 4, Piscataway, New Jersey, Nov./Dec. 1995, pp. 1942–1948.

[25] NS-3. *Network Simulator-3*. 2012. [Online]. Available: http://www.nsnam.org. [Accessed: Dec. 10, 2012].

Economics of Intelligent Selection of Wireless Access Networks in a Market-Based Framework: A Game-Theoretic Approach

Jakub Konka, James Irvine, and Robert Atkinson
Centre for Intelligent Dynamic Communications
University of Strathclyde
Glasgow G1 1XW, UK
Email: {jakub.konka, j.m.irvine, robert.atkinson}@strath.ac.uk

***Abstract*—The Digital Marketplace is a market-based framework where network operators offer communications services with competition at the call level. It strives to address a tussle between the actors involved in a heterogeneous wireless access network. However, as with any market-like institution, it is vital to analyze the Digital Marketplace from the strategic perspective to ensure that all shortcomings are removed prior to implementation. In this paper, we analyze the selling mechanism proposed in the Digital Marketplace. The mechanism is based on a procurement first-price sealed-bid auction where the network operators represent the sellers/bidders, and the end-user of a wireless service is the buyer. However, this auction format is somewhat unusual as the winning bid is a composition of both the network operator's monetary bid and their reputation rating. We create a simple economic model of the auction, and we show that it is mathematically intractable to derive the equilibrium bidding behavior when there are N network operators, and we make only generic assumptions about the structure of the bidding strategies. We then move on to consider a scenario with only two network operators, and assume that network operators use bidding strategies which are linear functions of their costs. This results in the derivation of the equilibrium bidding behavior in that scenario.**

***Keywords*-Wireless access networks; network selection; Digital Marketplace; economics; auction theory**

I. Introduction

This paper is an extension of the conference paper [1], and aims at providing a greater insight into the economics of intelligent network selection in the Digital Marketplace.

With the advent of 4th Generation wireless systems, such as WiMAX and 3GPP Long Term Evolution (LTE), the world of wireless and mobile communications is becoming increasingly diverse in terms of different wireless access technologies available [2], [3]; each of these technologies has its own distinct characteristics. Mirroring this diversity, multimode terminals (GSM/UMTS/Wi-Fi) currently dominate the market permitting the possibility of selecting the most appropriate access network to match the Quality of Service (QoS) requirements of a particular session/call. A number of approaches have examined this issue utilizing techniques as disparate as neural networks [4] and multiple attribute decision making [5]. The applicability of these techniques can be extended to fixed networks that employ multihoming where the problem becomes one of path selection [6], [7].

This work complements previous studies of intelligent network selection by considering economic aspects. From this perspective the exclusive one-to-one relationship between network operators and their subscribers no longer holds; subscribers are free to choose which operator and which access technology they would like to utilize at call set-up time. From the end-users' perspective, different coverage and QoS characteristics of each access network will lead to the ability to seamlessly connect at any time, at any place, and to the technology, which offers the best quality available for the best price. This is referred to as the *Always Best Connected* networking paradigm [8]. From the network operators' perspective, the integration of wireless access technologies will allow for more efficient usage of the network resources (by utilizing a wireless technology the most suitable to a particular service request), and may be the most economic way of providing both universal coverage and broadband access [2]. For example, a cellular network operator who also owns a set of Wi-Fi hot-spots will be able to offload the bandwidth intensive services from cellular base stations to Wi-Fi hot-spots. This should, in principle, reduce the potential cost to the network operator since instead of investing in additional cellular capacity, they can achieve the same (or better) results by investing into potentially cheaper Wi-Fi.

On the other hand, since many different actors with opposing interests are involved, it may also lead to a 'tussle' [9]. For example, the end-users seek to obtain the best quality for the best price, while the network operators are concerned with maximizing their profit and/or performing efficient load balancing. The conflict will become even more aggravated should the service provision be separated from the network operators [10]. Hence more sophisticated management techniques may be required to manage such a complex system.

In this paper, we analyze the network selection mechanism proposed in the Digital Marketplace (DMP) [11]. The DMP is a framework where network operators offer communications services with competition at the call level, and it

strives to address the tussle between the actors involved in a heterogeneous wireless access network. Within this framework, the network selection mechanism constitutes a sealed-bid auction. We create a simple economic model of the auction, and show that it is mathematically intractable to derive the equilibrium bidding behavior when there are N network operators competing in the DMP, and we make only generic assumptions about the structure of the bidding strategies. We then move on to consider a scenario with only two network operators, and assume that network operators use bidding strategies which are linear functions of their costs. This results in the derivation of the equilibrium bidding behavior in that scenario. The main goal of this research is to demonstrate and analyze the boundary conditions for such a market to function in the future. In this context, the participants could be cellular network operators or, alternatively, localized Wi-Fi hotspot operators competing for business.

The rest of this paper is organized as follows. In Section II, a brief summary of related work by other authors is given, while in Section III, an overview of the DMP is provided. Section IV presents the results of the analysis. Section V discusses future work, while Section VI draws conclusions.

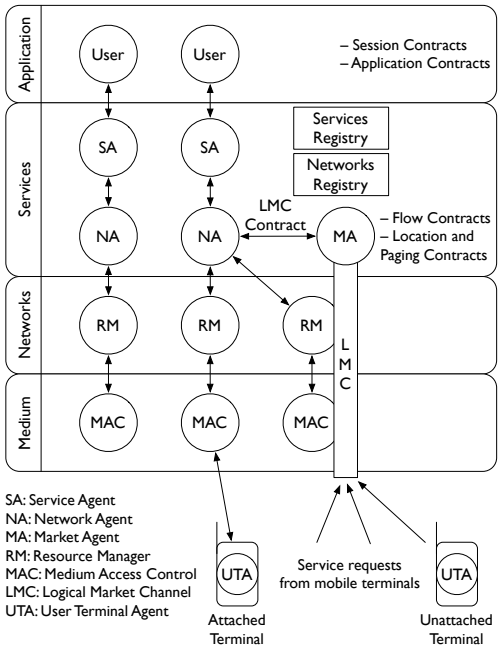


Figure 1. The Digital Marketplace (adapted from [11])

II. Related Work

Over the last decade, several different approaches have been proposed as possible solutions to the problem where economic competition is considered. Antoniou *et al.*, and Charilas *et al.* model the problem as a noncooperative game between wireless access networks, which aims at obtaining the best possible tradeoff between networks' efficiency and available capacity, while, at the same time, satisfying the end-users' QoS [12], [13]. Ormond *et al.* propose an algorithm for intelligent cost-oriented and performance-aware network selection, which maximizes consumer surplus [14], [15]. Niyato *et al.* propose two game-theoretic algorithms for intelligent network selection mechanism, which performs intelligent load balancing to avoid network congestion and performance degradation [16]. Khan *et al.* model the problem as a procurement second-price sealed-bid auction where network operators are the bidders and the end-user is the buyer [17], [18]. Lastly, Irvine *et al.* propose a market-based framework called the DMP, where network operators offer communications services with competition at the call level [11], [19], [20].

Although each proposed solution is technically valid, only the DMP strives to address tussle between the actors involved. Not only does the DMP consider the technical challenges but also the economic issues. However, as with any market-like institution, it is vital to analyze the DMP from the strategic perspective (using game theory, or otherwise) to ensure that all shortcomings are removed prior to implementation. This paper presents results of such an analysis.

III. The Digital Marketplace

The DMP was developed with the heterogeneous mobile and wireless communications environment in mind, where the end-users have the ability to select a network operator that reflects their preferences best on a per-call basis. In other words, the end-users have the freedom of choice, while the network operators manage service requests appropriately.

The conceptual framework of the DMP is shown in Figure 1. The DMP is defined using a four-layer communications stack: *application layer*, *services layer*, *networks layer*, and *medium layer*. The end-users who effectively reside in the application layer are able to negotiate network access on a per-call basis. To this end, they have two ways of accomplishing it: they can either go into a business relationship with a service provider (service agent, SA, in Figure 1) who will act on their behalf, or they can personally participate in the negotiation process with a network operator (network agent, NA). In both cases, the process is supervised by a market provider (market agent, MA), and takes place in the services layer. Before the negotiation occurs, the end-user is required to forward their service requirements to either the SA or the NA. This is done using a common communications channel referred to as a logical market channel (LMC). The LMC itself is negotiated between the MA and the registered NAs at the marketplace initialization stage.

The network selection mechanism in the DMP is based on a procurement first-price sealed-bid (FPA) auction. The network operators represent the sellers/bidders who compete for the right to sell their product (bearer service) to the end-user. However, unlike in a standard procurement FPA auction, here, network operators do not bid only on prices, but also on reputation; i.e., when selecting the winner, the end-user takes into consideration both the offered price of the product and the network operator's reputation. The reputation is directly proportional to the number of calls that have been decommitted in the past by the respective network operator. Since the network selection is intended to be performed on a per call basis, in a wireless environment, an important factor to consider while selecting an access network is the link/connection quality. There exists an extensive research base in the literature discussing technical constraints of the network selection problem (for example, see [21] for a survey of approaches); however, a few consider economic aspects. In this research, we suggest that poor link/connection quality strongly implies poor reputation.

Out of sealed-bid and sequential-bid auctions (such as English or Dutch auctions), an FPA auction was chosen as a selection mechanism due to the following reasons. Firstly, given the timing constraints in the DMP (e.g., the waiting time of the end-user for the call to be admitted), and the difficulty in predicting the number of bids placed until the winner is selected in a sequential-bid auction, sealed-bid auctions are deemed as the most appropriate [11]. Secondly, the rules governing a second-price sealed-bid auction may appear as counter-intuitive to the end-user; that is, the lowest bid secures the auction but the price paid equals the second-lowest bid [22]. Lastly, since the end-users not only base their network selection strategy on the offered price, but also on reputation, an FPA auction is the best fit to such a requirement.

An FPA auction, in an economic terminology, is an example of an allocation mechanism; that is, a system where economic transactions take place and goods are allocated [23]. As briefly mentioned in the Introduction, it is vital to analyze it from the strategic perspective, and establish what the most probable outcome will be; how the network operators will most likely bid; etc. In this way, all the shortcomings and inefficiencies can be addressed prior to implementation.

IV. Modeling and Analysis

The following notation and concepts are assumed throughout the rest of this paper.

1) Probability Theory and Statistics: Let X denote a random variable (r.v.) with the support $[a, b]$, where $a < b$ and $a, b \in \mathbb{R}$. By F_X we mean a cumulative distribution function of the X r.v.; therefore, for any $x \in \mathbb{R}$, $F_X(x) = P\{X \leq x\}$, where $P\{X \leq x\}$ denotes the probability of the event such that $X \leq x$. If F_X admits a density function, it shall be denoted by $f_X \equiv F'_X$.

The expected value of X, denoted by $E[X]$, is defined as $E[X] = \int_{-\infty}^{\infty} x dF_X(x)$. Similarly, if u is a function of X, then the expected value of $u(X)$ is defined as $E[u(X)] = \int_{-\infty}^{\infty} u(x) dF_X(x)$.

Let $X_1, \ldots, X_n$ be independent continuous r.v.s with distribution function F and density function $f \equiv F'$. If we let $X_{i:n}$ denote the ith smallest of these r.v.s, then $X_{1:n}, \ldots, X_{n:n}$ are called the *order statistics* [24], [25]. In the event that the r.v.s are independently and identically distributed (i.i.d.), the distribution of $X_{i:n}$ is

$$F_{X_{i:n}}(x) = \sum_{k=i}^{n} \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}, \quad (1)$$

while the density of $X_{i:n}$ can be obtained by differentiating Equation (1) with respect to x [26]. Hence,

$$f_{X_{i:n}}(x) = \frac{n!}{(n-i)!(i-1)!} f(x)(F(x))^{i-1}(1 - F(x))^{n-i}.$$

2) Game Theory: Let $\Gamma^B = [N, \{S_i\}, \{u_i\}, \Theta, F]$ be a *Bayesian game with incomplete information*. Formally, in this type of games, each player $i \in N$ has a utility function $u_i(s_i, s_{-i}, \theta_i)$, where $s_i \in S_i$ denotes player i's action, $s_{-i} \in S_{-i} = \times_{j \neq i} S_j$ denotes actions of all other players different from i, and $\theta_i \in \Theta_i$ represents the type of player i. Letting $\Theta = \times_{i \in N} \Theta_i$, the joint probability distribution of the $\theta \in \Theta$ is given by $F(\theta)$, which is assumed to be common knowledge among the players [27]–[29].

In game Γ^B, a *pure strategy* for player i is a function $\psi_i : \Theta_i \to S_i$, where for each type $\theta_i \in \Theta_i$, $\psi_i(\theta_i)$ specifies the action from the feasible set S_i that type θ_i would choose. Therefore, player i's pure strategy set Ψ_i is the set of all such functions.

Player i's *expected utility* given a profile of pure strategies $(\psi_1, \ldots, \psi_N)$ is given by

$$\tilde{u}_i(\psi_1, \ldots, \psi_N) = E[u_i(\psi_1(\theta_1), \ldots, \psi_N(\theta_N), \theta_i)], \quad (2)$$

where the expectation is taken over the realizations of the players' types, $\theta \in \Theta$. Now, in game Γ^B, a strategy profile $(\psi_1^*, \ldots, \psi_N^*)$ is a *pure-strategy Bayesian Nash equilibrium* if it constitutes a Nash equilibrium of game $\Gamma^N = [N, \{\Psi_i\}, \{\tilde{u}_i\}]$; that is, if for each player $i \in N$,

$$\tilde{u}_i(\psi_i^*, \psi_{-i}^*) \geq \tilde{u}_i(\psi_i, \psi_{-i}^*) \quad (3)$$

for all $\psi_i \in \Psi_i$, where $\tilde{u}_i(\psi_i, \psi_{-i})$ is defined as in Equation (2).

3) Incentive Compatibility, Individual Rationality and the Revelation Principle: Let $(\mathbf{Q}, \mathbf{M})$ be a direct mechanism where $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_{|N|})$ is an allocation rule, and $\mathbf{M} = (M_1, M_2, \ldots, M_{|N|})$ a payment rule. Let, as before, Θ_i be the set of all types of player i. The allocation rule Q_i for each player $i \in N$ is then defined as $Q_i : \Theta_i \to \Delta_i$ where Δ_i is the set of all probability distributions over Θ_i. Similarly,

the payment rule M_i for each player $i \in N$ is defined as $M_i : \Theta_i \to \mathbb{R}$ [22], [30].

A direct mechanism $(\mathbf{Q}, \mathbf{M})$ is said to satisfy *incentive compatibility* (IC) constraint if for all $i \in N$, for all $\theta_i \in \Theta_i$, and for all $\hat{\theta}_i \in \Theta_i$,

$$\tilde{\tilde{u}}_i(\theta_i) \equiv q_i(\theta_i)\theta_i - m_i(\theta_i) \geq q_i(\hat{\theta}_i)\theta_i - m_i(\hat{\theta}_i),$$

where

$$q_i(\hat{\theta}_i) = E[Q_i(\hat{\theta}_i, \theta_{-i})],$$

and

$$m_i(\hat{\theta}_i) = E[M_i(\hat{\theta}_i, \theta_{-i})].$$

In both cases, the expectation is taken over the realizations of all but player i types, $\theta_{-i} \in \Theta_{-i}$.

A direct mechanism $(\mathbf{Q}, \mathbf{M})$ is said to satisfy *individual rationality* (IR) constraint if for all $i \in N$, and for all $\theta_i \in \Theta_i$,

$$\tilde{\tilde{u}}_i(\theta_i) \geq 0.$$

In the paper, we will also make use of the very powerful Revelation Principle theorem [22], [31], [32]:

Theorem 1 (Revelation Principle)**.** *Given a mechanism and an equilibrium for that mechanism, there exists a direct mechanism in which (1) it is an equilibrium for each buyer to report his or her value truthfully and (2) the outcomes are the same as in the given equilibrium of the original mechanism.*

A. Problem Definition and Assumptions

The formal description of the network selection mechanism employed in the DMP is as follows. The model is a modified version of procurement FPA auction. Thus, formally, it represents a Bayesian game of incomplete information, Γ^B, as defined in Section IV-2. There are N network operators who bid for the right to sell their product to the end-user. With some abuse of notation, we will write N to denote the cardinality of the set N unless it becomes ambiguous where we will succumb to the standard notation of $|N|$.

Let $\beta : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$, defined by

$$\beta(b_i, r_i) = w_{price} \cdot b_i + w_{penalty} \cdot r_i \quad \text{for all } i \in N, \qquad (4)$$

denote the *compound bid*. Each network operator i is characterized by the utility function u_i such that

$$u_i(b, c, r) = \begin{cases} b_i - c_i & \text{if } \beta(b_i, r_i) < \min\limits_{j \neq i} \beta(b_j, r_j), \\ 0 & \text{if } \beta(b_i, r_i) > \min\limits_{j \neq i} \beta(b_j, r_j), \end{cases} \qquad (5)$$

where $b = (b_i, b_{-i})$ represents the monetary bid (or offered price) vector, $c = (c_i, c_{-i})$ the type vector, and $r = (r_i, r_{-i})$ the reputation rating vector. The type of each network operator is assumed to represent the cost of (or the minimum price for) the service under consideration. The winner of the auction is determined as the network operator whose compound bid is the lowest one; i.e., network operator i is the winner if

$$\beta(b_i, r_i) < \min_{j \neq i} \beta(b_j, r_j).$$

In the event that there is a tie

$$\beta(b_i, r_i) = \min_{j \neq i} \beta(b_j, r_j),$$

the winner is randomly selected with equal probability.

It is, moreover, assumed that the price and reputation weights $(w_{price}, w_{penalty})$ are announced by the end-user to all network operators before the auction. Thus, there is no uncertainty in knowing how much the end-user values the offered price of the service over the reputation of the network operator (or vice versa). Furthermore,

$$w_{price} + w_{penalty} = 1, \quad 0 \leq w_{price}, w_{penalty} \leq 1.$$

In order to simplify the notation, it is assumed throughout the rest of this paper that $w = w_{price}$. This reduces the definition of the compound bid in Equation (4) to

$$\beta(b_i, r_i) = wb_i + (1 - w)r_i \quad \text{for all } i \in N.$$

The set of network operators, N, is finite and the network operators are risk neutral. Furthermore, the end-user is risk neutral and does not have any budget constraints; that is, the end-user is prepared to accept any offer from the network operator.

The costs c_i for each network operator i are private knowledge. Thus, they are particular realizations of the r.v.s C_i for each i. Furthermore, it is assumed that each C_i is i.i.d. over the interval $[0, 1]$, and admits a continuous distribution function F_C and its associated density function f_C.

The reputation ratings r_i for each network operator $i \in N$ are common knowledge. It is assumed that each $r_i \in [0, 1]$ such that the higher the reputation, the lower the rating r_i. In earlier work [1], it was assumed that ratings are private knowledge. However, after analysis, it was concluded that this would contradict its purpose. The reputation of each network operator, in order to be meaningful, must be freely available to everyone, including the competitors of the network operators. For example, in the Amazon.com Marketplace, the buyers have the right to rate the seller they buy from on a scale from one to five (with five being the best), and these ratings are publicly available [33]. Similarly, on eBay, the buyers can leave sellers feedback (negative, neutral, or positive), which over time is viewed as reputation, and is also publicly available [34].

The bidding strategy functions $b_i : [0, 1] \to \mathbb{R}_+$ are nonnegative in value for all $i \in N$. The aim is to solve the game for pure-strategy Bayesian Nash equilibrium(-a) as defined in Equation (3), Section IV-2.

The problem will be divided into two cases: generic and restricted case. In the former, no additional assumptions

about the game than those already stated in the previous section will be made, and we will concentrate on finding a symmetric equilibrium. In the latter, on the other hand, the problem will be simplified by considering only two network operators, letting the costs be drawn from the uniform distribution, and focusing on bidding strategies, which are linear functions of cost.

B. Generic Case

Suppose that all network operators use the same strictly increasing in c_i bidding strategy function; i.e., $b_i = b_i(c_i) = b(c_i)$ for all $i \in N$. In this case, the equilibrium profile $(b^*, \ldots, b^*)$ is called *symmetric*. In its generic form, the problem proves complicated enough for the analytical solution not to be achievable using the existing methods of solving auctions. It would seem that since the problem is a modified version of the standard FPA, the standard analytical approach, found for example in [22], [35]–[37], should apply. However, this is not the case. To see why, note that each network operator i faces an optimization problem

$$\max_{b_i} E\left[b_i - c_i \,\middle|\, wb_i + (1-w)r_i < \min_{j\neq i}(wb(C_j) + (1-w)r_j)\right].$$

Noting that

$$\min_{j\neq i}(wb(C_j) + (1-w)r_j) \geq w\min_{j\neq i} b(C_j) + (1-w)\min_{j\neq i} r_j,$$

and assuming that $w \neq 0$, yields

$$\max_{b_i} E\left[b_i - c_i \,\middle|\, b^{-1}\left(b_i + \frac{1-w}{w}(r_i - \min_{j\neq i} r_j)\right) < \min_{j\neq i} C_j\right] \quad (6)$$

where we have used the fact that b is strictly increasing, and hence, it is invertible and $\min_x b(x) = b(\min_x x)$ for all x.

Let $C_{1:N-1} = \min_{j\neq i} C_j$ be the lowest order statistic of an i.i.d. random sample C_j for all $j \neq i$ with the distribution function $F_{C_{1:N-1}}$. Hence, the identity (6) becomes

$$\max_{b_i}\left(b_i - c_i\right)\left(1 - F_C\left(b^{-1}\left(b_i + \frac{1-w}{w}(r_i - \min_{j\neq i} r_j)\right)\right)\right)^{N-1} \quad (7)$$

where we have used the fact that the distribution function of an i^{th} order statistic of an i.i.d. random sample is defined as in Equation (1).

Finally, recalling that at a symmetric equilibrium $b_i = b(c_i)$ and letting $k = \frac{(1-w)}{w}(r_i - \min_{j\neq i} r_j)$, the identity (7) becomes

$$\begin{aligned}&\frac{d}{dc_i} b\left(b^{-1}(b(c_i)+k)\right)\cdot\left[1 - F_C(b^{-1}(b(c_i)+k))\right]^{N-1}\\ &= (N-1)(b(c_i) - c_i)\left[1 - F_C(b^{-1}(b(c_i)+k))\right]^{N-2}\\ &\cdot f_C(b^{-1}(b(c_i)+k)).\end{aligned} \quad (8)$$

It is rather difficult (if even possible) to solve the resulting ordinary differential equation in (8). Therefore, it can be concluded that even serious simplification of the problem is not enough to heuristically derive an optimal bidding strategy function for each network operator i.

However, it is possible to gain some insight into the problem by analyzing a handful of boundary (or special) cases; that is, $w = 0$, $w = 1$, and $r_i = r_j$ for all $i \neq j$. In all three cases, the problem simplifies enough for the analytical analysis to be tractable, as presented below.

1) Special Case $w = 0$: When $w = 0$, the utility function simplifies to

$$u_i(b, c, r) = \begin{cases} b_i - c_i & \text{if } r_i < \min\limits_{j\neq i} r_j, \\ 0 & \text{if } r_i > \min\limits_{j\neq i} r_j. \end{cases} \quad (9)$$

Since the reputation ratings, r_i, are common knowledge, the probability of winning, i.e., the probability of the event such that $r_i < \min_{j\neq i} r_j$ for all i, is either 0 or 1, and does not depend on the value of the bid, b_i. In other words, each network operator knows in advance whether they won, tied, or lost based on their own and their opponents reputation ratings since these are deterministic in nature. Hence, it is clear that the network operator with the lowest reputation rating will have an incentive to bid abnormally high since they are guaranteed a win regardless of the value of their bid. The remaining network operators, on the other hand, will be indifferent to the value of the submitted bids as it is impossible for them to win regardless of the values of their bids. In case of a tie, i.e., in case there is more than one network operator with the lowest reputation rating, each has an equal probability of winning the auction, and this probability is independent of the values of their bids. Hence, in this case, the network operators also have an incentive to bid abnormally high. Formally,

Proposition 1. *Suppose c_i is i.i.d. over the interval $[0,1]$ for all $i \in N$ and $r_i \in [0,1]$ for all $i \in N$ is common knowledge. Let $N_0 \subseteq N$ be the set of all those network operators with the lowest reputation rating. If $w = 0$, then every network operator $j \in N_0$ will have an incentive to bid abnormally high, i.e., $b_j \to \infty$, while every remaining network operator $k \in N \setminus N_0$ will be indifferent to the value of their bid.*

The formal proof of Proposition 1 is given in Appendix A.

In real life, the end-user will be constrained by a fixed budget. Therefore, when $w = 0$, the real value of the bid will not tend to infinity; rather it is expected to oscillate in the region of the highest price the end-user is willing to pay for the service. In this way, the network operator will extract the entire consumer surplus from the end-user who is looking for a premium service of the best possible quality.

2) Special Case $w = 1$: When $w = 1$, on the other hand, the problem reduces to that of standard FPA auction. The utility of each network operator i becomes

$$u_i(b, c, r) = \begin{cases} b_i - c_i & \text{if } b_i < \min\limits_{j\neq i} b_j, \\ 0 & \text{if } b_i > \min\limits_{j\neq i} b_j. \end{cases} \quad (10)$$

Network operator i, conjecturing that other network operators follow the symmetric equilibrium bidding strategy b and

submit their costs truthfully, solves

$$\begin{aligned}
&\max_{b_i} E\left[b_i - c_i \,\middle|\, b_i < \min_{j\neq i} b(C_j)\right] \\
&= \max_{b_i} E\left[b_i - c_i \,\middle|\, b^{-1}(b_i) < \min_{j\neq i} C_j\right] \\
&= \max_{b_i} E\left[b_i - c_i \,\middle|\, b^{-1}(b_i) < C_{1:N-1}\right] \\
&= \max_{b_i} \int_{b^{-1}(b_i)}^{1} (b_i - c_i) dF_{C_{1:N-1}}(t) \\
&= \max_{b_i} (b_i - c_i)(1 - F_{C_{1:N-1}}(b^{-1}(b_i))), \qquad (11)
\end{aligned}$$

where, as before, $C_{1:N-1} = \min_{j\neq i} C_j$ is the lowest order statistic of an i.i.d. random sample C_j for all $j \neq i$ with the distribution function $F_{C_{1:N-1}}$, and its associated density $f_{C_{1:N-1}}$. The first-order condition yields

$$1 - F_{C_{1:N-1}}(b^{-1}(b_i)) - (b_i - c_i)\frac{f_{C_{1:N-1}}(b^{-1}(b_i))}{\frac{d}{db_i} b(b^{-1}(b_i))} = 0. \qquad (12)$$

Recalling that at a symmetric equilibrium $b_i = b(c_i)$, the identity (12) becomes

$$\frac{d}{dc_i} b(c_i) - b(c_i)\frac{f_{C_{1:N-1}}(c_i)}{1 - F_{C_{1:N-1}}(c_i)} = -c_i \frac{f_{C_{1:N-1}}(c_i)}{1 - F_{C_{1:N-1}}(c_i)}.$$

Since $b(1) = 1$, we have

$$\begin{aligned}
b(c_i) &= \frac{1}{1 - F_{C_{1:N-1}}(c_i)} \int_{c_i}^{1} t dF_{C_{1:N-1}}(t) \\
&= \frac{N-1}{(1 - F_C(c_i))^{N-1}} \int_{c_i}^{1} t(1 - F_C(t))^{N-2} f_C(t) dt. \qquad (13)
\end{aligned}$$

The symmetric bidding strategy in Equation (13) constitutes a symmetric pure-strategy Bayesian Nash equilibrium of the standard FPA auction when $w = 1$. Formally,

Proposition 2. *Suppose c_i is i.i.d. over the interval $[0,1]$ for all $i \in N$ and $r_i \in [0,1]$ for all $i \in N$ is common knowledge. If $w = 1$, then the symmetric equilibrium bidding strategy function of the standard procurement first-price sealed-bid auction,*

$$b^*_{FPA}(c_i) = \frac{1}{1 - F_{C_{1:N-1}}(c_i)} \int_{c_i}^{1} t dF_{C_{1:N-1}}(t), \qquad (14)$$

constitutes a symmetric pure-strategy Bayesian Nash equilibrium of the Digital Marketplace variant of a procurement first-price sealed-bid auction.

The formal proof of Proposition 2 can be found in [1].

The next natural question to ask is whether b^*_{FPA} constitutes an equilibrium for $w \neq 1$. The following conjecture summarizes this point,

Conjecture 3. *Suppose c_i are i.i.d. over the interval $[0,1]$ for all $i \in N$ and $r_i \in [0,1]$ for all $i \in N$ are common knowledge. If the symmetric equilibrium bidding strategy function of the standard procurement first-price sealed-bid auction, b^*_{FPA}, constitutes a symmetric pure-strategy Bayesian Nash equilibrium of the Digital Marketplace variant of a procurement first-price sealed-bid auction, then $w = 1$.*

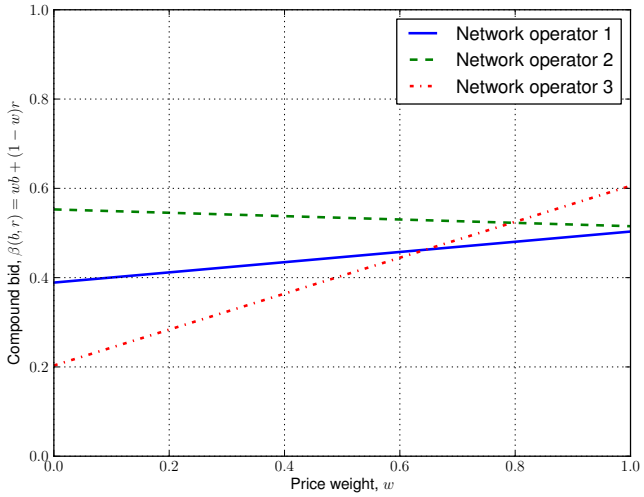


Figure 2. The performance of standard FPA bidding strategy, b^*, for costs, reputation ratings and bids aggregated in Table I

The conjecture can be rephrased as "If $w \neq 1$, then b^*_{FPA} does not constitute a symmetric pure-strategy Bayesian Nash equilibrium of the Digital Marketplace variant of a procurement first-price sealed-bid auction." The formal proof of this statement is rather difficult. However, the following argument shows why it might hold.

Suppose for the time being that $b^*(c_i) = b^*_{FPA}(c_i)$ for every value of the price weight $w \in [0,1]$. It is possible to estimate numerically how well such a bidding strategy performs for all values of w. To this end, a simple Monte Carlo simulation scenario was constructed where the network operators' costs and reputation ratings were pseudo-randomly generated and drawn from a uniform distribution $\mathcal{U}[0,1]$.

Table I and Figure 2 depict the output from a single simulation for $N = 3$ network operators. In this particular example, for $w \in (0.65, 1]$, network operator 1 who is characterized by the lowest cost of all three network operators, wins the auction; that is, his compound bid is the lowest. At $w = 0.65$, an intersection occurs of network operator 1's and 3's compound bids, and after that, for $w \in [0, 0.65)$, network operator 3 becomes the winner. If the simulation was repeated n times, and the intersection would fall within a close neighborhood of $w = 0.65$ in the vast majority of cases, then b^* is quite likely to be an equilibrium bidding strategy in the interval $w \in (0.65, 1]$. This is predicated on the fact that, as $w \to 1$, the offered price dominates the value of the compound bid; that is, the offered price is weighted more than the reputation rating (see Equation (4)).

The methodology is as follows:

1) Generate cost/reputation rating/bid triplet using the Monte Carlo methods.
2) Find the winner for $w = 1$, network operator i, say (in Figure 2 that would be network operator 1).
3) Decrease the value of w until network operator i no

Table I
THE OUTPUT FROM A SINGLE RUN OF THE MONTE CARLO SIMULATION FOR $N = 3$ NETWORK OPERATORS

	Cost, c_i	**Reputation rating**, r_i	**Bid**, $b^*(c_i)$
Network operator 1	0.2548	0.3889	0.5032
Network operator 2	0.2728	0.5528	0.5152
Network operator 3	0.4084	0.2031	0.6056

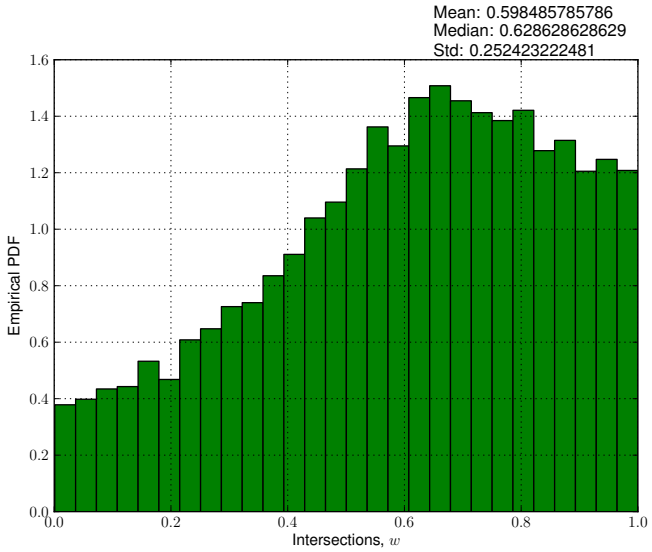


Figure 3. The histogram of intersections, simulated for $n = 10,000$ runs and $N = 3$ network operators

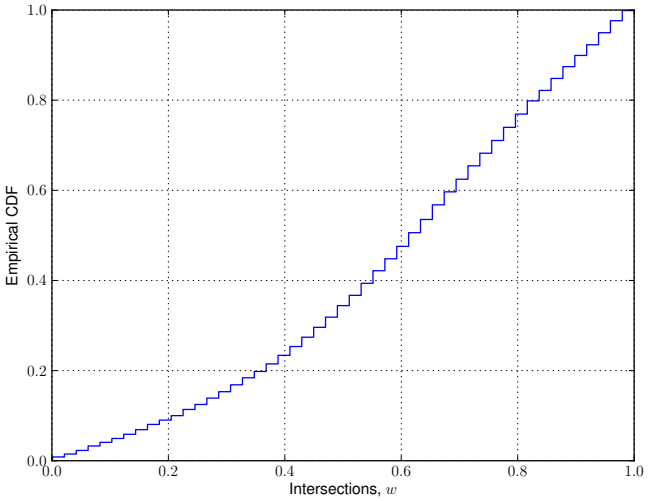


Figure 4. The empirical probability distribution associated with the histogram in Figure 3

longer wins, and save the value of w for which that happens. Henceforth, such an event shall be denoted by I, and called *the event when an intersection has occurred.*

4) If the intersection did not occur, $I = 0$, increase the counter that counts the frequency of such an event, and then discard that run.
5) Repeat n number of times.

By way of example, Figure 3 depicts the empirical density function of the intersections simulated for $n = 10,000$ runs and $N = 3$ network operators, while Figure 4 shows the associated empirical distribution function. The probability of an intersection occurring equals $P\{I = 1\} = 0.67$. It can be concluded from the figures that, on average, the intersections occur at $\bar{w} \approx 0.6$, which represents the mean of the distribution. However, the peak observed in a close neighborhood of $\bar{w}$ is not significant enough to conclude that bidding according to b^* is the best strategy one can take for $w \in (\bar{w}, 1]$.

A more formal argument goes as follows. Figure 4 depicts the probability that an intersection has occurred within an interval $(-\infty, w]$ given that an intersection has occurred, $I = 1$; that is, if the former event is denoted by W, then the figure describes $P\{W \in (-\infty, w] \mid I = 1\}$. From this, the probability of winning for network operator i (as defined in the list above) given any w is

$$\begin{aligned} &P\{\text{winning} \mid w\} = \\ &= 1 - P\{W \in [w, \infty) \cap I = 1\} \\ &= 1 - P\{W \in [w, \infty) \mid I = 1\} P\{I = 1\} \\ &= 1 - (1 - P\{W \in (-\infty, w] \mid I = 1\})\, P\{I = 1\}. \end{aligned} \tag{15}$$

In order to verify Equation (15), set $w \in \{0.25, 0.75\}$ and run a Monte Carlo simulation, which counts the number of times when the network operator with the lowest cost is the winner; i.e., the winner of the auction for $w = 1$. When $w = 0.25$,

$$P\{\text{winning} \mid w = 0.25\} = 1 - (1 - 0.13)0.67 = 0.4171$$

according to Equation (15), while the numerically obtained result equals

$$P\{\text{winning} \mid w = 0.25\} = 0.4136.$$

When $w = 0.75$,

$$P\{\text{winning} \mid w = 0.75\} = 1 - (1 - 0.68)0.67 = 0.7856$$

according to Equation (15), while the numerically obtained result equals

$$P\{\text{winning} \mid w = 0.75\} = 0.7866.$$

Clearly, the prediction based on Equation (15) converges to the numerically obtained result. Moreover, it is worth noting that for $w = 0.25$, bidding according to b^* guarantees the probability of winning for the network operator with the lowest cost of only 0.4171, which is below 50%. Thus, the network operators will definitely deviate from b^* for low values of w. On the other hand, for $w = 0.75$, b^* seems to achieve a relatively high probability of winning for the network operator with the lowest cost; i.e., the probability of 0.7856. However, the argument is incomplete in the sense that it only considers the probability of winning rather than the expected utility.

3) Special Case $r_i = r_j$: In the last extreme case, when all the network operators are characterized by the same reputation rating, i.e., when $r_i = r_j$ for all $i \neq j$ and $w \neq 0$, it can easily be verified that the problem simplifies to the special case $w = 1$. To see why, let $r = r_i$ for all $i \in N$. Then, for all $i \in N$ and $w \neq 0$

$$\begin{aligned} & \beta(b_i, r) < \min_{j \neq i} \beta(b_j, r) \\ \iff & \frac{1}{w}\left(b_i + \frac{1-w}{w} r\right) < \frac{1}{w} \min_{j \neq i}\left(b_j + \frac{1-w}{w} r\right) \\ \iff & b_i + \frac{1-w}{w} r < \min_{j \neq i} b_j + \frac{1-w}{w} r \\ \iff & b_i < \min_{j \neq i} b_j. \end{aligned}$$

Hence, the utility of each network operator i simplifies to

$$u_i(b, c, r) = \begin{cases} b_i - c_i & \text{if } b_i < \min\limits_{j \neq i} b_j, \\ 0 & \text{if } b_i > \min\limits_{j \neq i} b_j. \end{cases}$$

Formally,

Corollary 4. *Suppose c_i is i.i.d. over the interval $[0, 1]$ for all $i \in N$ and $r_i \in [0, 1]$ for all $i \in N$ is common knowledge. Suppose $r_i = r_j$ for all $i \neq j$, and $w \neq 0$. Then, the problem simplifies to the special case $w = 1$, and hence, b^*_{FPA} is the symmetric equilibrium bidding strategy (Proposition 2).*

C. Restricted Case $N = 2$

In this section, we will restrict our attention to only two network operators. Since the problem in its generic form proved too complex to be solved analytically, this section will explore whether in a much simplified scenario it is possible to find a closed-form solution. From the mathematical standpoint, restricting the number of network operators to two considerably simplifies the optimization problem that each network operator faces, since it is no longer necessary to consider the minimum of β in the specification of network operators' utility function (Equation (5)).

To this end, let $N = 2$. The utility function for each network operator i thus becomes

$$u_i(b, c, r) = \begin{cases} b_i - c_i & \text{if } \beta(b_i, r_i) < \beta(b_j, r_j), \\ \frac{1}{2}(b_i - c_i) & \text{if } \beta(b_i, r_i) = \beta(b_j, r_j), \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

Furthermore, the assumption about the symmetric equilibrium profile is relaxed; that is, network operators are permitted to use differing bidding strategies.

The analysis is conducted in two steps. Firstly, it is assumed that information is complete; that is, that each network operator not only knows their own cost and reputation rating, but also those of their opponent's. Secondly, the standard case is considered; that is, that the reputation ratings of the network operators are assumed to be known, while the costs are private knowledge.

1) Complete Information: Here, we assume that information is complete; i.e., that each network operator knows their own and their opponent's cost and reputation rating. In total, there are 7 different bidding scenarios to consider.

Figure 5 shows the first 4 cases for which $r_i < r_j$. (Notice that exactly the same reasoning applies to the situation when $r_i > r_j$.) If $c_i < c_j$, network operator i is guaranteed a victory and a positive profit as long as they bid within the highlighted part of the $\beta(b, r)$ curve depicted in Figure 5a. Thus, their optimal bidding strategy would be to bid slightly less than their opponent's compound bid evaluated at their opponent's cost, $\beta(c_j, r_j)$; that is, $b_i = c_j + \frac{1-w}{w}(r_j - r_i) - \epsilon$ where $\epsilon > 0$ is very small. Network operator j, on the other hand, should find it optimal to bid $b_j = c_j$. To see why, suppose network operator j bids $\hat{b}_j > c_j$. Since network operator i's reputation and cost are strictly lower than those of network operator j's, they can undercut the network operator j's bid by a small amount so that $\hat{b}_i < \hat{b}_j$ and still make positive profit. But, in response, network operator j will find it optimal to lower their bid so that it undercuts that of network operator i's; that is, $\hat{b}_j < \hat{b}_i$. This process will continue until one of the network operators is forced to bid their cost. Since network operator i's reputation and cost are strictly lower than those of network operator j's, we conclude that $b_j = c_j$ and $b_i = c_j + \frac{1-w}{w}(r_j - r_i) - \epsilon$ where $\epsilon > 0$ is very small.

If $c_i = c_j$, arguing in the similar manner as previously, network operator i's optimal bidding strategy would be to bid $b_i = c_j + \frac{1-w}{w}(r_j - r_i) - \epsilon$ where $\epsilon > 0$ is very small; while network operator j should bid $b_j = c_j$ (Figure 5b).

If $c_i > c_j$, there are two cases to consider. If $\beta(c_i, r_i) < \beta(c_j, r_j)$, then network operator i still has some room for maneuver, and should find it optimal to bid $b_i = c_j + \frac{1-w}{w}(r_j - r_i) - \epsilon$ where $\epsilon > 0$ is very small; while network operator j to bid $b_j = c_j$ (Figure 5c). If $\beta(c_i, r_i) \geq \beta(c_j, r_j)$, on the other hand, the roles are reversed, and network operator j should find it optimal to bid $b_j = c_i + \frac{1-w}{w}(r_i - r_j) - \epsilon$ where $\epsilon > 0$ is very small; while network operator i to bid $b_i = c_i$ (Figure 5d).

Figure 6 depicts the remaining 3 cases for which $r_i = r_j$. If $c_i < c_j$, network operator i's optimal bidding strategy would be to bid $b_i = c_j - \epsilon$ where $\epsilon > 0$ is very small; while network operator j should bid $b_j = c_j$ (Figure 6a).

If $c_i = c_j$, both network operators should bid their costs;

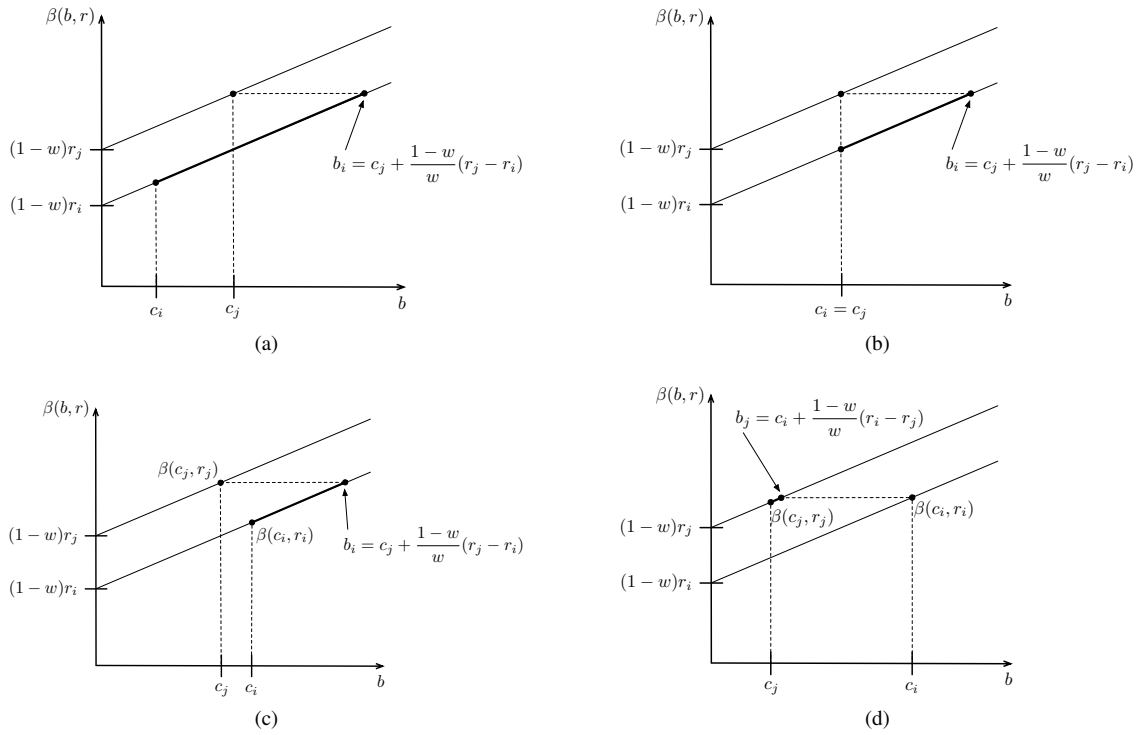


Figure 5. Different bidding scenarios for $r_i < r_j$: (a) $c_i < c_j$, (b) $c_i = c_j$, (c) $c_i > c_j$ with $\beta(c_i, r_i) < \beta(c_j, r_j)$, and (d) $c_i > c_j$ with $\beta(c_i, r_i) \geq \beta(c_j, r_j)$

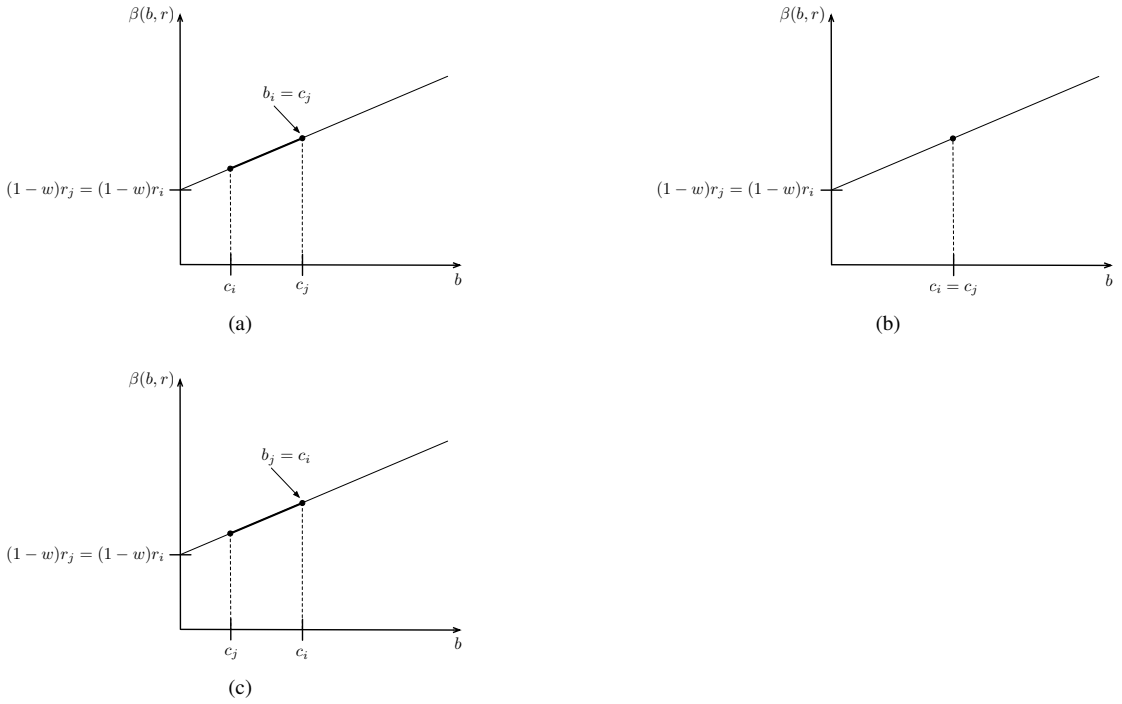


Figure 6. Different bidding scenarios for $r_i = r_j$: (a) $c_i < c_j$, (b) $c_i = c_j$, and (c) $c_i > c_j$

that is, $b_i = c_i$ and $b_j = c_j$ (Figure 6b).

If $c_i > c_j$, network operator j's optimal bidding strategy would be to bid $b_j = c_i - \epsilon$ where $\epsilon > 0$ is very small; while network operator i should bid $b_i = c_i$ (Figure 6c).

It can be concluded that the bidding strategies depend only on costs if $r_i = r_j$. In the remaining cases, they are asymmetric in the sense that the winning network operator is characterized by

$$b_i = c_j + \frac{1-w}{w}(r_j - r_i) - \epsilon \quad \text{with } \epsilon > 0 \text{ being very small,}$$

while the losing network operator by bidding their own cost

$$b_j = c_j.$$

Hence, when dealing with incomplete information, we will exploit these results by concentrating on equilibrium bidding strategies, which are linear functions of cost.

2) Incomplete Information: Here, we assume the standard case; that is, that reputation ratings for both network operators are known at the time of bidding; however, their costs are private knowledge. Suppose that the network operators use a strategy function $b_i : [0,1] \to \mathbb{R}$ defined by the rule

$$b_i(c_i) = m_i + n_i c_i, \quad \text{for all } m_i \in \mathbb{R}, n_i > 0, \tag{17}$$

and the costs are independently drawn from the uniform distribution over the interval $[0,1]$. In other words, (although somewhat counter-intuitive) we allow for negative bids from the network operators. The motivation for such an assumption will become clear later on in this section. Notice, moreover, that the strategy function is assumed to be linear in cost. Each network operator i faces an optimization problem

$$\max_{b_i} E\left[b_i - c_i \mid wb_i + (1-w)r_i < w(m_j + n_j C_j) + (1-w)r_j\right]. \tag{18}$$

If $w = 0$, then the result described in Proposition 1, Section IV-B1, holds. Otherwise, for $0 < w \leq 1$, each network operator i solves

$$\begin{aligned}
&\max_{b_i} E\left[b_i - c_i \,\middle|\, \frac{1}{n_j}\left(b_i + \frac{1-w}{w}(r_i - r_j) - m_j\right) < C_j\right] \\
&= \max_{b_i} \int_{\frac{1}{n_j}(b_i + \frac{1-w}{w}(r_i - r_j) - m_j)}^{1} (b_i - c_i) dF_C(t) \\
&= \max_{b_i} \left(b_i - c_i\right)\left(1 - \frac{1}{n_j}b_i - \frac{1}{n_j}\left(\frac{1-w}{w}(r_i - r_j) - m_j\right)\right).
\end{aligned} \tag{19}$$

The first-order condition yields

$$\begin{aligned}
&1 - \frac{2}{n_j}b_i + \frac{1}{n_j}c_i - \frac{1}{n_j}\left(\frac{1-w}{w}(r_i - r_j) - m_j\right) = 0 \\
\iff\ & b_i = \frac{n_j}{2} - \frac{1}{2}\left(\frac{1-w}{w}(r_i - r_j) - m_j\right) + \frac{1}{2}c_i.
\end{aligned} \tag{20}$$

(Notice that the second-order condition is satisfied; i.e., $\frac{d^2}{db_i^2}E[\cdot|\cdot] = -\frac{2}{n_j} < 0$ since $n_j > 0$.) Similar argument for network operator j yields

$$b_j = \frac{n_i}{2} - \frac{1}{2}\left(\frac{1-w}{w}(r_j - r_i) - m_i\right) + \frac{1}{2}c_j. \tag{21}$$

Table II
AN EXEMPLARY SET OF COST-REPUTATION PAIRS FOR TWO NETWORK OPERATORS

	Cost, c_i	**Reputation rating**, r_i
Network operator 1	0.75	0.25
Network operator 2	0.25	0.75

Thus, it follows

$$\begin{cases}
n_i &= n_j = \frac{1}{2}, \\
m_i &= \frac{n_j}{2} - \frac{1}{2}\left(\frac{1-w}{w}(r_i - r_j) - m_j\right), \\
m_j &= \frac{n_i}{2} - \frac{1}{2}\left(\frac{1-w}{w}(r_j - r_i) - m_i\right).
\end{cases}$$

Solving the above equations simultaneously yields the equilibrium bidding strategy, for all i

$$b_i'(c_i) = \frac{1}{2} - \frac{1-w}{3w}(r_i - r_j) + \frac{1}{2}c_i.$$

Formally,

Proposition 5. *Let there be $N = 2$ network operators. Suppose c_i is independently drawn from uniform distribution over the interval $[0,1]$ for all $i \in N$, and $r_i \in [0,1]$ for all $i \in N$ is common knowledge. Then the equilibrium bidding strategy for all $w \in (0,1]$ is given by*

$$b_i'(c_i) = \frac{1}{2} - \frac{1-w}{3w}(r_i - r_j) + \frac{1}{2}c_i. \tag{22}$$

The formal proof of Proposition 5 is given in Appendix A. Observe that the pair of strategies (b_i', b_j') does not constitute a symmetric equilibrium.

By way of example, Table II depicts a particular set of cost-reputation pairs of two network operators. Figure 7 shows the value of the compound bid, β, for different values of w for both network operators, while Figure 8 depicts the value of the bid (or offered price), b_i', for different values of w for both network operators. The numerical data in Table II suggests that network operator 2 should be the winner for the values of $w \to 1$ since network operator 2's cost is strictly lower than that of their opponent's. On the other hand, network operator 1 should be the winner for the values of $w \to 0$ since network operator 1's reputation rating is strictly lower that that of their opponent's (which implies that network operator 1's reputation is strictly higher than that of their opponent's). This prediction agrees with the numerical output shown in Figures 7 and 8. Let w_c denote the value of w for which an intersection between the compound bids of both network operators occurs (if it exists). In Figure 7, $w_c = 0.4$. Hence, network operator 2 wins the auction for the values of $w \in (w_c, 1]$, while network operator 1 for the values of $w \in [0, w_c)$. Notice, moreover, that since the range of the strategy function, b_i, was modified to span the entire real line, that is,

$$b_i : [0,1] \to \mathbb{R},$$

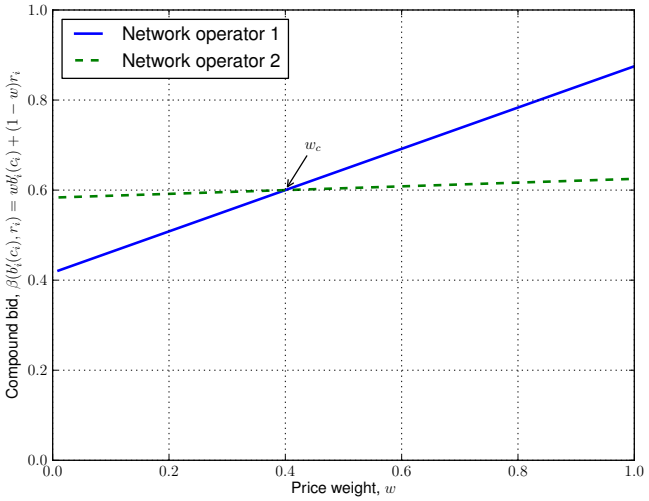


Figure 7. Compound bid plotted against the price weight

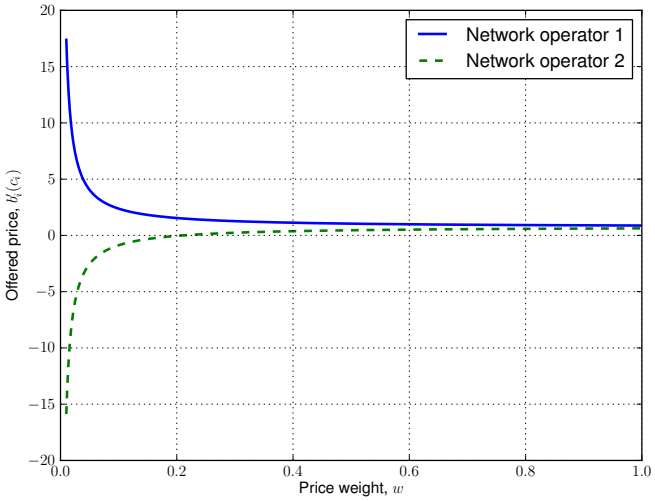


Figure 8. Offered prices (bids) plotted against the price weight

network operator 2, as a result, bids below their cost for values of $w < w_c$ (Figure 8). However, this does not automatically disqualify the equilibrium bidding strategies given by Equation (22). The following observations show why.

Firstly,

Proposition 6. *Suppose both network operators bid according to b'_i bidding strategies. Then they are guaranteed nonnegative profit in case of winning (or a tie).*

The formal proof of Proposition 6 is given in Appendix A.

The proposition implies that even though the equilibrium bidding strategies suggest that one of the network operators may bid negatively, they will not win the auction, and hence, are guaranteed profit at worst equal to zero. Therefore, the possibility of one of the network operators bidding below their cost or negatively will not matter to any of the network operators, and will not lead to an outcome in which the service is sold for a negative price.

Secondly, let $(\mathbf{Q}, \mathbf{M})$ be the direct mechanism induced by the equilibrium bidding strategies, b'_i, in Equation (22) where $\mathbf{Q} = (Q_i, Q_j)$ and $\mathbf{M} = (M_i, M_j)$. Here, Q_i represents the allocation rule defined by

$$Q_i(c_i, c_j) = \begin{cases} 1 & \text{if } \beta(b'_i(c_i), r_i) < \beta(b'_j(c_j), r_j), \\ \frac{1}{2} & \text{if } \beta(b'_i(c_i), r_i) = \beta(b'_j(c_j), r_j), \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

while M_i is the payment rule defined by

$$M_i(c_i, c_j) = Q_i(c_i, c_j) b'_i(c_i). \tag{24}$$

Suppose network operator j reveals their cost truthfully. The equilibrium payoff function for network operator i characterized by cost c_i but revealing $\hat{c}_i$ is

$$\begin{aligned} \tilde{\tilde{u}}_i(\hat{c}_i) &= E\left[M_i(\hat{c}_i, C_j) - c_i Q_i(\hat{c}_i, C_j)\right] \\ &= E\left[(b'_i(\hat{c}_i) - c_i) Q_i(\hat{c}_i, C_j)\right] \\ &= E\left[b'_i(\hat{c}_i) - c_i \mid \beta(b'_i(\hat{c}_i), r_i) < \beta(b'_j(C_j), r_j)\right]. \end{aligned} \tag{25}$$

It turns out that it is in network operator i's best interest to reveal their cost truthfully as well; i.e., $\hat{c}_i = c_i$. Moreover, both network operators cannot be better off by not participating in the auction; i.e., their equilibrium payoff function is nonnegative, $\tilde{\tilde{u}}_i(c_i) \geq 0$. Formally,

Proposition 7. *The direct mechanism $(\mathbf{Q}, \mathbf{M})$ where $\mathbf{Q} = (Q_i, Q_j)$ and $\mathbf{M} = (M_i, M_j)$ (with Q_i and M_i defined in Equations (23) and (24) respectively) satisfies both the IC and IR constraints.*

Thirdly, suppose that economic agents are computers who bid on behalf of the network operators. This assumption is reasonable since there currently are estimated 6.1 billion mobile subscribers around the world [38]. In other words, bidding on a per-call basis would have to be automated by the network operators in order to make the process manageable. One way of achieving such an automation would be to utilize the concept of a direct mechanism. In a direct mechanism, economic agents submit their costs (which need not be truthful) directly to the mechanism, which then computes the bids and chooses the winner on their behalf. By the Revelation Principle (which is stated in Section IV-3), we know that for every mechanism and an equilibrium for that mechanism, there exists an incentive compatible direct mechanism, which yields the same outcomes as in the given equilibrium of the original mechanism. In our case, the direct mechanism $(\mathbf{Q}, \mathbf{M})$ is the direct representation of the DMP variant of an FPA. Since it is incentive compatible, it is in best interest of the economic

agents to reveal their costs truthfully. Furthermore, because it is individually rational, it is also in their best interest to participate in the mechanism [22]. Therefore, the possibility of one of the network operators bidding below their cost or negatively will not matter to any of the network operators and will not lead to an outcome in which the service is sold for a negative price.

V. Future Work

In the restricted case, the possibility of one of the network operators bidding below their cost or negatively might seem counter-intuitive and irrational. Therefore, one of the future directions will include an in-depth analysis of this problem. The most straightforward solution is to constrain the optimization problem in Equation (18); that is, each network operator i tries to solve

$$\max_{b_i} E\left[b_i - c_i \mid wb_i + (1-w)r_i < w(m_j + n_j C_j) + (1-w)r_j\right]$$
$$\text{subject to} \quad c_i - b_i \leq 0.$$

The constraint $c_i - b_i \leq 0$ ensures that each network operator bids above or equal to their cost. However, this problem is much more complicated than its unconstrained version in Equation (18). Not only is it necessary to solve the nonlinear constrained optimization problem for each network operator i, but also it needs to be done simultaneously [39]. The preliminary analysis of the problem, which employs the application of the Karush-Kuhn-Tucker Conditions theorem seems to suggest that the most likely candidate for the solution would be

$$b_i^*(c_i) = \max\left\{c_i, b_i'(c_i)\right\} \quad \text{for all } i \in N.$$

However, this has yet to be verified.

VI. Conclusions

This paper has presented the results of the game-theoretic analysis of network selection mechanism proposed in the Digital Marketplace. All things considered, it can be concluded that the analytical analysis of the Digital Marketplace variant of procurement first-price sealed-bid auction is mathematically intractable for all but special cases considered in this paper. It is, however, vital to have at least partially accurate predictions of the behavior of the network operators prior to implementation.

In the generic case, where there are N network operators and costs are drawn from an arbitrary continuous distribution, derivation of the equilibrium bidding behavior is complicated. Nevertheless, some light was shed on the problem in a handful of special cases: $w = 0$, $w = 1$, and $r_i = r_j$. In the first case, we showed that network operators will find it beneficial to submit abnormally high bids, since their bid is independent of the probability of winning the auction. In the remaining two cases, when $w = 1$ and $r_i = r_j$, we showed that the problem reduces to a standard procurement first-price sealed-bid auction, and therefore, the symmetric equilibrium bidding behavior of the standard procurement first-price auction constitutes an equilibrium of the Digital Marketplace auction.

In the restricted case, where there are two network operators and costs are uniformly distributed, we successfully derived the equilibrium bidding strategies that are linear functions of cost. However, we showed that the derived bidding strategy functions constitute an asymmetric equilibrium; that is, their closed-form expression is not identical for both network operators. This implies that the analysis of the case with more than two network operators might not be analytically possible, and hence, indirectly explains the reason for unsuccessful analysis of the generic case. Furthermore, we showed that although the derived equilibrium bidding behavior allows for negative bids, it does not lead to negative profit in case of winning (or a tie) of either network operator. In fact, we established that the direct mechanism representation of the Digital Marketplace auction satisfies both individual rationality and incentive compatibility constraints. Therefore, if the auction were to be automated through the use of a direct mechanism, the network operators would find it in their best interest to participate in the auction, and they would reveal their costs truthfully.

Appendix
Proofs

Proof of Proposition 1: Let $w = 0$ and let $|N_0| = M$ be the number of network operators with the lowest reputation rating such that $M \in \mathbb{Z}_+$. Since N is finite and $N_0 \subset N$, then $M \leq |N|$. Now, each $j \in N_0$ is facing a maximization problem

$$\max_{b_j} \frac{1}{M}\left(b_j - c_j\right), \quad \text{for all } j \in N_0.$$

Since $1 \leq M \leq |N|$, and since $b_j \in \mathbb{R}_+$ and $\mathbb{R}_+$ is not bounded from above, this implies that the maximization problem is unbounded; that is, $b_j \to \infty$ for all $j \in N_0$.

The remaining network operators $k \in N - N_0$ will try to solve

$$\max_{b_k} 0, \quad \text{for all } k \in N - N_0,$$

since $r_k > r_j = \min_{i \in N} r_i$. Hence, each network operator $k \in N - N_0$ is indifferent to the value of their bid, which concludes the proof. ■

Proof of Proposition 5: Suppose there are two network operators: network operator 1 and network operator 2 with cost-reputation pairs (c_1, r_1) and (c_2, r_2) respectively. Suppose that network operator 2 follows b_2' equilibrium bidding strategy. We will argue that it is optimal for network operator 1 to follow b_1' equilibrium bidding strategy. First, notice that b_1' is strictly increasing and continuous function of cost (similarly is b_2'). Suppose that network operator 1 bids an amount b_1. Since b_1' is strictly increasing, there exists unique cost $\hat{c}_1$

such that $\hat{c}_1 = b_1'^{-1}(b_1)$. Network operator 1's expected utility when bidding $b_1'(\hat{c}_1)$ is

$$\begin{aligned}
&\tilde{u}_1(b_1'(\hat{c}_1), c_1) \\
&= E\left[b_1'(\hat{c}_1) - c_1 \mid wb_1'(c_1) + (1-w)r_1 < wb_2'(c_2) + (1-w)r_2\right] \\
&= \frac{1}{2}\left(1 - \frac{2}{3}\cdot\frac{1-w}{w}(r_1 - r_2) + \hat{c}_1 - 2c_1\right) \\
&\cdot \left(1 - \hat{c}_1 - \frac{2}{3}\cdot\frac{1-w}{w}(r_1 - r_2)\right).
\end{aligned}$$

We thus obtain that

$$\tilde{u}_1(b_1'(c_1), c_1) - \tilde{u}_1(b_1'(\hat{c}_1), c_1) = \frac{1}{2}(c_1 - \hat{c}_1)^2 \geq 0$$

regardless of whether $\hat{c}_1 \geq c_1$ or $\hat{c}_1 \leq c_1$. We have thus argued that if network operator 2 follows b_2', network operator 1 with a cost c_1 cannot benefit by bidding anything other than $b_1'(c_1)$. Similar argument can be used to show that it is optimal for network operator 2 to follow b_2' when network operator 1 is following b_1'. Hence, (b_1', b_2') constitutes a Bayesian-Nash equilibrium profile. Similar argument can be used to show that it is optimal for network operator 2 to reveal their cost truthfully, which concludes the proof. ■

Proof of Proposition 6: Let there be two network operators: network operator 1 and network operator 2 with cost-reputation pairs (c_1, r_1) and (c_2, r_2) respectively. Suppose that both network operators follow the equilibrium bidding strategy, b_i'. Without loss of generality, we need to show that network operator 1's bid is always at least as high as their cost whenever they win or draw with network operator 2; that is, $b_1'(c_1) \geq c_1$.

First of all, notice that if $r_1 \leq r_2$,

$$b_1'(c_1) = \frac{1}{2} - \frac{1-w}{3w}(r_1 - r_2) + \frac{1}{2}c_1 \geq \frac{1}{2}(1 + c_1) \geq c_1,$$

for all $c_1 \in [0, 1]$. Thus, we need only to consider the case when $r_1 > r_2$.

Suppose $r_1 > r_2$. If $c_1 > c_2$, and since $b_1'(c_1)$ is strictly increasing in c_1, network operator 1 will lose for all values of $w \in (0, 1]$. If $c_1 = c_2$, network operator 1 will lose for all values of $w \in (0, 1)$, except at $w = 1$ when there will be a draw. But at $w = 1$, network operator 1's bid is at least as high as their cost; i.e.,

$$b_1'(c_1) = \frac{1}{2}(1 + c_1) \geq c_1, \quad \text{for all } c_1 \in [0, 1].$$

If $c_1 < c_2$, it is sufficient to show that the intersection of $b_1'(c_1)$ and c_1 in terms of w can never occur before the intersection of $\beta(b_1'(c_1), r_1)$ and $\beta(b_2'(c_2), r_2)$. First of all, we need to check that both intersections do occur; that is,

$$b_1'(c_1) = c_1 \iff w = \frac{1}{1 + \frac{3}{2}\cdot\frac{1-c_1}{r_1 - r_2}}.$$

Similarly,

$$\beta(b_1'(c_1), r_1) = \beta(b_2'(c_2), r_2) \iff w = \frac{1}{1 + \frac{3}{2}\cdot\frac{c_2 - c_1}{r_1 - r_2}}.$$

Since $r_1 > r_2$ and $c_1 < c_2$, we have $0 < r_1 - r_2 \leq 1$ and $0 < c_2 - c_1 \leq 1$. Therefore, this implies

$$0 < w = \frac{1}{1 + \frac{3}{2}\cdot\frac{1-c_1}{r_1 - r_2}} \leq 1,$$

and

$$0 < w = \frac{1}{1 + \frac{3}{2}\cdot\frac{c_2 - c_1}{r_1 - r_2}} \leq 1.$$

Now, suppose that the intersection of $b_1'(c_1)$ and c_1 occurs before that of $\beta(b_1'(c_1), r_1)$ and $\beta(b_2'(c_2), r_2)$. We must thus have

$$\frac{1}{1 + \frac{3}{2}\cdot\frac{c_2 - c_1}{r_1 - r_2}} < \frac{1}{1 + \frac{3}{2}\cdot\frac{1-c_1}{r_1 - r_2}} \iff \frac{1 - c_2}{r_1 - r_2} < 0.$$

But since $c_2 \in [0, 1]$ and $r_1 > r_2$ by assumption,

$$0 < \frac{1 - c_2}{r_1 - r_2}$$

we reach a contradiction, and this concludes the proof. ■

Proof of Proposition 7: Let there be two network operators: network operator 1 and network operator 2 with cost-reputation pairs (c_1, r_1) and (c_2, r_2) respectively. Suppose that both network operators participate in the direct mechanism $(\mathbf{Q}, \mathbf{M})$. Firstly, we show that the mechanism is incentive compatible. Without loss of generality, suppose that network operator 2 truthfully submits their cost to the mechanism. We argue that it is optimal for network operator 1 to also submit their cost truthfully. Suppose to the contrary; that is, that network operator 1 has an incentive not to reveal their cost truthfully by submitting $\hat{c}_1$. Thus, their expected utility becomes

$$\begin{aligned}
&\tilde{\tilde{u}}_1(\hat{c}_1) \\
&= E\left[b_1'(\hat{c}_1) - c_1 \,\middle|\, 2b_1'(\hat{c}_1) - 1 + \frac{4}{3}\cdot\frac{1-w}{w}(r_1 - r_2) < C_2\right] \\
&= \left(\frac{1}{2} - \frac{1}{3}\cdot\frac{1-w}{w}(r_1 - r_2) + \frac{1}{2}\hat{c}_1 - c_1\right) \\
&\cdot \left(1 - \hat{c}_1 - \frac{2}{3}\cdot\frac{1-w}{w}(r_1 - r_2)\right).
\end{aligned}$$

The first-order condition yields $\hat{c}_1 = c_1$ and the second-order condition is satisfied. Hence, this shows that $(\mathbf{Q}, \mathbf{M})$ is incentive compatible.

Secondly, we show that $(\mathbf{Q}, \mathbf{M})$ is individually rational. Since the mechanism is incentive compatible, each network operator reveals their cost truthfully. Hence, for all c_1

$$\begin{aligned}
\tilde{\tilde{u}}_1(c_1) &= \left(\frac{1}{2} - \frac{1}{3}\cdot\frac{1-w}{w}(r_1 - r_2) - \frac{1}{2}c_1\right) \\
&\cdot \left(1 - c_1 - \frac{2}{3}\cdot\frac{1-w}{w}(r_1 - r_2)\right) \\
&= \frac{1}{2}\left(1 - c_1 - \frac{2}{3}\cdot\frac{1-w}{w}(r_1 - r_2)\right)^2 \geq 0.
\end{aligned}$$

Therefore, $(\mathbf{Q}, \mathbf{M})$ is individually rational. ■

ACKNOWLEDGMENTS

The authors would like to thank Dr. Alex Dickson from the Department of Economics at the University of Strathclyde for his invaluable comments and guidance through this research.

REFERENCES

[1] J. Konka, R. Atkinson, and J. Irvine, "Economic Aspects of Intelligent Network Selection: A Game-Theoretic Approach," in *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2011. UBICOMM '11. The Fifth International Conference on*, 20-25 Nov. 2011.

[2] R. Beaubrun, "Integration of Heterogeneous Wireless Access Networks," in *Heterogeneous Wireless Access Networks: Architectures and Protocols* (E. Hossain, ed.), ch. 1, pp. 1–18, Springer, 2009.

[3] P. TalebiFard, T. Wong, and V. C. M. Leung, "Integration of Heterogeneous Wireless Access Networks with IP-based Core Networks: The Path to Telco 2.0," in *Heterogeneous Wireless Access Networks: Architectures and Protocols* (E. Hossain, ed.), ch. 2, pp. 19–54, Springer, 2009.

[4] J. Espi, R. Atkinson, D. Harle, and I. Andonovic, "An Optimum Network Selection Solution for Multihomed Hosts Using Hopfield Networks," in *Networks (ICN), 2010 Ninth International Conference on*, pp. 249–254, April 2010.

[5] C. Shen, W. Du, R. Atkinson, J. Irvine, and D. Pesch, "A mobility framework to improve heterogeneous wireless network services," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 7, no. 1, pp. 60–69, 2011.

[6] Q. Wang, R. Atkinson, and J. Dunlop, "Design and Evaluation of Flow Handoff Signalling for Multihomed Mobile Nodes in Wireless Overlay Networks," *Elsevier Computer Networks*, vol. 52, no. 8, pp. 1647–1674, 2008.

[7] Q. Wang, T. Hopf, F. Filali, R. Atkinson, J. Dunlop, E. Robert, and L. Aginako, "QoS-Aware Network-Supported Architecture to Distribute Applications Flows Over Multiple Network Interfaces for B3G Users," *Springer Wireless Personal Communications*, vol. 48, no. 1, pp. 113–140, 2009.

[8] E. Gustafsson and A. Jonsson, "Always Best Connected," *Wireless Communications, IEEE*, vol. 10, no. 1, pp. 49–55, 2003.

[9] D. D. Clark and J. Wroclawski, "Tussle in Cyberspace: Defining Tomorrow's Internet," in *SIGCOMM'02*, (Pittsburgh, Pennsylvania, USA), 19–23 August 2002.

[10] J. Bush, J. Irvine, and J. Dunlop, "A Digital Marketplace for Tussle in Next Generation Wireless Networks," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, pp. 1–5, Sept. 2009.

[11] G. Le Bodic, D. Girma, J. Irvine, and J. Dunlop, "Dynamic 3G Network Selection for Increasing the Competition in the Mobile Communications Market," in *Vehicular Technology Conference, 2000. IEEE VTS-Fall VTC 2000. 52nd*, vol. 3, pp. 1064–1071, 2000.

[12] J. Antoniou and A. Pisillides, "4G Converged Environment: Modeling Network Selection as a Game," in *16th IST Mobile and Wireless Communication Summit*, (Budapest), July 2007.

[13] D. Charilas, O. Markaki, and E. Tragos, "A Theoretical Scheme for Applying Game Theory and Network Selection Mechanisms in Access Admission Control," in *Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on*, pp. 303–307, May 2008.

[14] O. Ormond, G.-M. Muntean, and J. Murphy, "Economic Model for Cost Effective Network Selection Strategy in Service Oriented Heterogeneous Wireless Network Environment," in *NOMS'06*, 2006.

[15] O. Ormond, G.-M. Muntean, and J. Murphy, "Evaluation of an Intelligent Utility-Based Strategy for Dynamic Wireless Network Selection," in *MMNS'06*, pp. 158–170, 2006.

[16] D. Niyato and E. Hossain, "Dynamics of Network Selection in Heterogeneous Wireless Networks: An Evolutionary Game Approach," *Vehicular Technology, IEEE Transactions on*, vol. 58, pp. 2008–2017, May 2009.

[17] M. A. Khan, U. Toseef, S. Marx, and C. Goerg, "Game-Theory Based User Centric Network Selection with Media Independent Handover Services and Flow Management," in *Communication Networks and Services Research Conference (CNSR), 2010 Eighth Annual*, pp. 248–255, May 2010.

[18] M. Khan, U. Toseef, S. Marx, and C. Goerg, "Auction-based Interface Selection with Media Independent Handover Services and Flow Management," in *Wireless Conference (EW), 2010 European*, pp. 429–436, April 2010.

[19] J. Irvine, C. McKeown, and J. Dunlop, "Managing Hybrid Mobile Radio Networks with the Digital Marketplace," in *Vehicular Technology Conference, 2001. VTC 2001 Fall. IEEE VTS 54th*, vol. 4, pp. 2542–2546, 2001.

[20] J. Irvine, "Adam Smith Goes Mobile: Managing Services Beyond 3G with the Digital Marketplace," in *Wireless Conference (EW), 2002 European. Invited paper to*, 2002.

[21] A. Sgora and D. Vergados, "Handoff Prioritization and Decision Schemes in Wireless Cellular Networks: a Survey," *Communications Surveys Tutorials, IEEE*, vol. 11, pp. 57–77, quarter 2009.

[22] V. Krishna, *Auction Theory*. Academic Press, second ed., 2010.

[23] Compiled by the Prize Committee of the Royal Swedish Academy of Sciences, "Scientific background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2007: Mechanism Design Theory." [Online]. Available: http://www.academia.edu/156730/http_nobelprize.org_nobel_prizes_economics_laureates_2007_ecoadv07.pdf [Accessed Dec. 4, 2012].

[24] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics*. Classics in Applied Mathematics 54, SIAM, 2008.

[25] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley & Sons, third ed., 2003.

[26] S. M. Ross, *Introduction to Probability Models*. Elsevier, tenth ed., 2010.

[27] R. B. Myerson, *Game Theory: Analysis of Conflict*. Harvard University Press, 1997.

[28] R. Gibbons, *A Primer in Game Theory*. Financial Times/Prentice Hall, 1992.

[29] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. Oxford University Press, 1995.

[30] R. B. Myerson, "Optimal Auction Design," *Mathematics of Operations Research*, vol. 6, pp. 58–73, Feb. 1981.

[31] R. B. Myerson, "Incentive Compatibility and the Bargaining Problem," *Econometrica*, vol. 47, pp. 61–73, Jan. 1979.

[32] M. Harris and A. Raviv, "Allocation Mechanisms and the Design of Auctions," *Econometrica*, vol. 49, pp. 1477–1499, Nov. 1981.

[33] Amazon.com Inc., "Amazon.com Help: Rating Your Amazon Marketplace Seller." [Online]. Available: http://www.amazon.com/gp/help/customer/display.html/ref=hp_rel_topic?ie=UTF8&nodeId=537806 [Accessed Dec. 4, 2012].

[34] eBay Inc., "eBay.com: All about Feedback." [Online]. Available: http://pages.ebay.com/help/feedback/allaboutfeedback.html [Accessed Dec. 4, 2012].

[35] R. P. McAfee and J. McMillan, "Auctions and Bidding," *Journal of Economic Literature*, vol. 25, pp. 699–738, Jun. 1987.

[36] R. G. Hansen, "Auctions with Endogenous Quantity," *RAND Journal of Economics*, vol. 19, no. 1, pp. 44–58, 1988.

[37] K. G. Dastidar, "On Procurement Auctions with Fixed Budgets," *Research in Economics*, vol. 62, pp. 72–91, June 2008.

[38] Ericsson, "Traffic and Market Data Report," tech. rep., Nov. 2011.

[39] C. Griffin, "Game Theory: Penn State Math 486 Lecture Notes." [Online]. Available: http://www.personal.psu.edu/cxg286/Math486.pdf [Accessed Dec. 4, 2012].

Location-Aware Routing for Service-Oriented Opportunistic Computing

Nicolas Le Sommer and Yves Mahéo
IRISA, Université de Bretagne-Sud, France
{Nicolas.Le-Sommer,Yves.Maheo}@univ-ubs.fr

***Abstract*—Smartphones, tablets, netbooks and laptops are intensively used every day by a large part of the population. These devices—which are equipped with Wi-Fi interfaces—can form disconnected mobile ad hoc networks (DMANETs) dynamically. These networks may allow service providers, such as local authorities, to deliver new kinds of services in a wide area (e.g., a city) without resorting to the infrastructure-based networks of mobile phone operators. This paper[1] presents OLFServ, a new location-aware forwarding protocol dedicated to service-oriented opportunistic computing in DMANETs. This protocol implements several self-pruning heuristics allowing mobile nodes to decide whether they efficiently contribute in the message delivery. The protocol has been implemented in a service-oriented middleware platform, and has been validated through simulations, which proved its efficiency.**

Keywords*—*Opportunistic Computing*; *Mobile Ad hoc Networks

I. Introduction

Over the last years, handheld devices such as smartphones or tablets have become widely spread and used through the population. These devices, which are equipped with wireless communication interfaces—often complemented by GPS (Global Positioning System) receivers and various sensors—, allow their users to connect to the Internet and to use services hosted in remote servers just as if they were at home using a wired connection. This kind of service provision knows a great development, but it relies on a fixed and often heavy infrastructure, and is not without constraints for the client when considering for instance the cost of resorting to licensed frequency bands (Universal Mobile Telecommunications System, General Packet Radio Service) or the limited geographical scope of a Wi-Fi hotspot.

An alternative has been envisioned since several years through mobile ad hoc networking. Mobile handheld devices can form mobile ad hoc networks spontaneously, and this ability can be exploited in order to artificially extend networks composed of some sparsely distributed infostations with a view to offering a wide service access to end-users. An illustration of this kind of network is shown in Figure 1: devices with Wi-Fi interfaces operated in ad hoc mode are present in the environment; most of them are hold by mobile users and few of them, the infostations, are fixed. In practice, because of the potentially low density of devices, their mobility and the short communication range of wireless interfaces, the topology of such networks suffers from frequent and unpredictable changes. The network is regularly fragmented in several distinct communication islands thus entailing an intermittent connectivity between devices and the impossibility to ensure an end-to-end connectivity. For these reasons, this type of network is called a DMANET for Disconnected Mobile Ad hoc Network.

In DMANETs, devices can communicate directly only when they are in range of one another. Intermediate nodes can be used to relay a message from a source to its destination following the "store, carry and forward" principle. The routes are therefore computed dynamically at each hop while the messages are forwarded towards their destination(s). Each node receiving a message for a given destination is thus expected to transmit a copy of the message to one or several of its neighbors. When no forwarding opportunity exists (e.g., no other nodes are in the transmission range, or the neighbors are evaluated as not suitable for that communication) the node stores the message and waits for future contact opportunities with other devices to forward the message. Thanks to this principle, a message can be delivered even if the client and the destination are not present simultaneously in the network, or if they are not in the same network island at emission time.

Devising an efficient routing based on the "store, carry, and forward" principle has been the subject of many research efforts in the so-called domain of Opportunistic Networking [2]. The main problem is to establish a compromise between the speed at which the message reaches its destination and the resources consumed globally in the network, namely the storage space required in the intermediate devices and the bandwidth used when transmitting messages between devices. Flooding the network with copies of the message is known to be the fastest way to attain the destination in theory but its cost is considered prohibitive. On the other hand, keeping a single copy of the message in the network and passing it from one device to its neighbor when possible is an economical solution but tends to slow down—if not jeopardize—the propagation of the message towards its destination. A common approach is to allow an intermediate device to generate a limited number of copies of the message and leverage on contextual information for selecting the best devices to which these copies are conveyed. The considered context can take various forms, related for instance to records of encounters with other devices or to device's location.

Although routing is a key aspect in DMANETs, it should not

[1] This paper is an extended version of a previous description of our work [1]. It gives a more detailed explanation of the rationale and mechanics of the proposed protocol, as well as complementary experimentation results.

Figure 1. Illustration of a disconnected MANET formed by infostations and devices carried by people strolling in a city.

be considered as the ultimate objective but rather as a first step towards middleware tools adapted to distributed application development. Indeed, legacy applications (often based on strong connectivity assumptions) cannot be straightforwardly transposed into the specific context of DMANETs, or do not take full benefit of the pervasive aspect of DMANETs. The effective emergence of new applications is dependent on the capacity to discover, compose and exploit heterogeneous resources spread on a disconnected network. The notion of Opportunistic Computing has been introduced to emphasize the gap between issues related to opportunistic networking, that mainly aims at forwarding message packets, and those related to application design and implementation [3], [4]. Because of its intrinsic loosely-coupled nature well adapted to opportunistic computing, a first obvious paradigm to investigate is service provisioning: hardware or software resources available in the network are abstracted as services. A service is hosted by a device that plays the role of service provider. Other devices in the network, acting as clients, will try to discover provided services so as to be able to invoke them remotely. An intermediate selection phase may take place before invocation, when the client is able to choose between several services. Service provisioning in connected networks has been extensively explored (Web Services are a well-known example) but in the framework of DMANETs, issues regarding discovery, selection and invocation introduced by this paradigm are seldom addressed[2]. The case in which services are provided only by fixed infostations is particularly interesting because the range of services susceptible to be deployed on this kind of platforms is very large compared with what can be done on mobile devices. Indeed, infostations are stable, not as constrained as mobile devices in terms of resources (primarily regarding power), and their potential connection to the Internet allows an easy access to a huge amount of information.

This paper presents OLFServ, a new opportunistic and location-aware forwarding protocol we have designed in order to support both service discovery and service invocation in DMANETs. OLFServ is a key element of a middleware platform we develop to investigate service provisioning in DMANETs [5]. Based on the location data collected by the platform from the wireless interface and/or the GPS receiver of the device, OLFServ makes it possible to perform an efficient and geographically-based broadcast of both service advertisements and service discovery requests, as well as a location-driven service invocation. OLFServ implements several self-pruning heuristics allowing intermediates nodes to decide themselves if they are "good" relays to deliver the messages they receive from their neighbors (i.e., if they contribute to bring a message closer to its destination). These heuristics aim to

- progressively refine the area where a message can be disseminated until reaching its destination;
- perform source routing when it is possible;
- support the client mobility by computing the area where the client is expected to be when it receives its response;
- avoid message collisions by implementing a backoff mechanism.

Thanks to these heuristics, only a small subset of relevant intermediate nodes will forward the messages in given geographical areas or in given directions.

The remainder of the paper is organized as follows. Section II brings to the fore the main issues that must be addressed in order to discover and to deliver some services in DMANETs efficiently. Section III presents the assumptions on which protocol OLFServ is based, the detailed specifications of the self-pruning heuristics it implements, and how it works on an example. Section IV presents some simulations results we obtained for OLFServ. Research works dealing with routing protocols in DMANETs are presented in Section V. Section VI summarizes our contribution.

II. Rationale for Service-Oriented Opportunistic Computing

When targeting DMANETs, the service-oriented opportunistic computing paradigm introduces new issues compared to the mere provision of message passing. These issues pertain namely to the discovery, the selection and the invocation of pervasive services, imposing de facto the design of new routing protocols suited to both discovery and delivery of services, as well as the development of middleware platforms supporting distributed computing tasks in environments where disconnections and network partitions are the rule. These aspects are discussed in the remainder of this section.

A. Service discovery

In disconnected, or partially connected, MANET, no device is stable enough, or accessible permanently, to act as a service registry. Mobile clients should therefore be responsible for discovering the services offered in the network reactively and/or proactively, and for maintaining their own list of services. The reactive discovery is usually achieved by processing the unsolicited service advertisements broadcast by service providers, while the proactive discovery is performed by broadcasting

[2]To our knowledge, except in our previous work, service provisioning in opportunistic networks has been specifically studied only by the European SCAMPI project (http://www.ict-scampi.eu).

service discovery requests in the network and by processing the advertisements returned in response by providers. In such a distributed discovery process, all mobile nodes receiving an advertisement or a discovery request are not expected to rebroadcast this message systematically and immediately, because if they do so, they will generate too much network traffic and could even lead to network congestion. To cope with this problem, which is known as the broadcast storm problem [6], some heuristics must be devised in order to reduce the number of senders and to broadcast the messages asynchronously. Moreover, based on the "store, carry and forward" principle, the discovery messages can be disseminated in a wide area, even if the services are relevant only in a restricted one. Thus, it seems to be suitable to circumscribe the dissemination of these messages geographically, as well as to limit their dissemination in the network by defining a life time and a maximum number of hops.

B. Service selection

A selection process may precede the invocation, when the opportunity is given to the client application to choose among several service providers. Thus, it could be interesting to select a provider according to its location, and to transparently select another one among a set of relevant ones when the current provider becomes inaccessible. In previous works, we proposed two different solutions for this issue: one that relies on a content-based service invocation [7] and another one that relies on a dynamic and transparent update of the service references [5]. These two solutions have been implemented in the service management layer of our middleware platform.

C. Service invocation

In opportunistic networks, no end-to-end routes are maintained between a client and a provider by an underlying dynamic routing protocol such as AODV or OLSR. A priori, a node does not know which is the best next forwarder among its neighbors for reaching the destination. In order to avoid a "blind" message forwarding, some solutions have been proposed over the last years [8], [9], [10], [11], [12], [13]. These solutions mainly rely on the computation of a delivery probability based on contextual properties [12], on a history of contacts [10], or on both [13], [9]. Nevertheless, these solutions often consider that nodes move following regular mobility patterns, and that their future (direct or indirect) encounters can be predicted. Computing such an history and a prediction is a tricky problem, especially in an environment where people often stroll and move randomly such as in a city, questioning de facto such assumptions. Moreover, during the invocation process, such probabilities must be computed twice: once in order to deliver the invocation request to the service provider, and another time to deliver the response to the client. Indeed, the client and the intermediate nodes are likely to move during this process, the forwarding path followed by the response can therefore be different from that taken by the request.

In order to increase the message delivery ratio and to reduce the delivery time, several copies of a message are usually generated in the network. In order not to process a request or a response several times, such a redundancy should be hidden from both the client applications and the software services, and be controlled by the routing protocol itself. Moreover, a mobile node should stop forwarding a request for which it has already received a response.

Opportunistic communications introduce a certain delay in the service discovery and invocation processes. Although client applications must be able to tolerate this delay and to deal with extended disconnection periods, it is suitable to devise solutions that provide end-users with a certain quality of service in term of responsiveness. Consequently, the protocol should not implement a purely periodic and proactive message emission, but instead should adopt a reactive behavior as far as possible. It should be sensitive to events such as the arrival of a new neighbor, the reception of a new message or the location changes.

Finally, like the service discovery messages, the service invocation requests and the service responses must be circumscribed to the area where the service must be offered. Both a lifetime and a maximum number of hops must also be assigned to these messages in order to reduce their propagation.

III. The OLFServ protocol

In the remainder of this section, we present OLFServ, an opportunistic and location-aware forwarding protocol we have designed so as to address the issues identified in the previous section. OLFServ aims at supporting the discovery and the invocation of software services in DMANETs such as those formed by fixed infostations and handheld devices used by nomadic people. It implements an efficient and geographically-constrained broadcast of both service advertisements and service discovery requests, as well as a location-driven forwarding of service invocation requests and service responses. OLFServ is a key element of an OSGi service-oriented middleware platform we have developed in order to support service provision in "challenging" pervasive environments. This platform provides some facilities in order to compute the location of a mobile node according to the coordinates generated by the embedded GPS receiver or to the Wi-Fi signal and the location properties exhibited by the neighbor nodes as presented in [5].

A. Assumptions

The OLFServ protocol relies on 3 main assumptions:

1) Both mobile hosts and fixed infostations are aware of their geographical location and able to compare their location with that of another host. Mobile hosts are expected to indicate their destination/direction if they know them.
2) Mobile hosts are able to perceive their one-hop neighborhood. This neighborhood is obtained using specific messages (beacons) sent by each node periodically.
3) Each mobile host is able to temporarily store the messages it receives, and can associate to each of them some pieces of information, and especially the IDs of the nodes that are known to have received them.

B. Overview of the protocol

1) Heurisitics:

OLFServ is an event-driven protocol that implements self-pruning heuristics. The originality of this protocol resides in the adaptation of several well known heuristics to the context of service provisioning in DMANETs, and their combination in a coherent platform. The main implemented heuristics are the following:

Contention resolution in message forwarding: Like DFCN (Delayed Flooding with Cumulative Neighborhood) [14], which proposes a bandwith-efficient broadcast algorithm for MANETs, OLFServ introduces a backoff mechanism in order to avoid message collisions at message reforwarding time. From this point of view, a node is expected to compute a forwarding delay for each message it receives, and to forward messages when their delay expires. Moreover, a node will abstain from forwarding a message if it perceives that all of its neighbors have already received it (the message was forwarded by at least one of its neighbors before it forwards the message itself, and its one-hop neighborhood is a subset of the set of nodes that are expected to have received the message yet). In addition, in OLFServ, this forwarding delay has two components: one that is inversely proportional to the distance from the last forwarder and another one that is a random value (used in the backoff mechanism). Therefore, only the farther nodes are likely to forward a message, thus improving the geographical propagation of messages while reducing the number of emissions.

Geographically-driven message forwarding: At each step, a message will be forwarded only by the nodes closer to the destination.

Content-based message forwarding: Mobile nodes can establish some correlations between the discovery requests and the advertisements, as well as between the invocation requests and the responses. Thanks to this heuristic, a mobile node receiving an invocation request is expected to send back to the client the response it previously stored for this request instead of forwarding it towards its destination, obviously if this one is still valid.

Source routing forwarding: Nodes can estimate if a message was forwarded quickly (i.e., if a message was relayed following an end-to-end path), and to perform source routing if so. OLFServ is thus able to exploit end-to-end routes when they exist, reducing the propagation time and the number of message copies. If the source routing failed, because an intermediate node becomes unreachable, the selective and controlled broadcast is used. These last two heuristics aims at improving the quality of service offered to end-users in term of responsiveness.

2) Events:

In OLFServ, five kinds of events are considered:

- the reception of a message;
- the expiration of the forwarding delay associated with a message;
- the location changes;
- the arrival of a new neighbor;
- the failure in the source routing process.

The first and the last events induce a reactive behavior of the protocol regarding the message forwarding, whereas the other events induce a proactive behavior.

Before giving a detailed specification of the OLFServ protocol, let's see how the above-mentioned heuristics operate in both the service discovery process and the service invocation phase. From this point of view, let us consider the disconnected MANET depicted in Figure 2, which will, for the sake of illustration, be composed of a set of mobile devices carried by pedestrians and a fixed infostation I that offers a service that is relevant only in the geographical area represented by the dotted rectangle. Moreover, let's suppose that one of these mobile hosts, namely node C, is interested in the service proposed by I. The network, which is currently composed of the six distinct communication islands shown in Figure 2, is expected to evolve in an unpredictable manner according to the nodes' mobility. Nevertheless, in order to illustrate our purposes, we will consider subsequently that node C and node N_6 follow the materialized paths so as to reach different destinations at times t_1, t_2, t_3 and t_4.

a) Service discovery: The invocation of a remote service is conditioned by the preliminary discovery of this service. Consequently, in order to call the service offered by I, node C must discover this service. For the sake of illustration, let us consider that infostation I has injected in the network an advertisement A including its location, the geographical area where the service can be accessed, a date of emission, a lifetime, a maximum number of hops this advertisement is allowed to make, and the set of nodes that are expected to receive this advertisement (i.e., I, N_1, N_2, N_3, N_4 and N_5). Nodes N_1, N_2, N_3, N_4 and N_5, which will receive message A first, will store this message locally and will compute a forwarding delay in order not to rebroadcast message A simultaneously.

The coverage radio area of a node is partitioned in several concentric rings. The forwarding delay algorithm (see Algorithm 2) allows mobile nodes located approximately at the same distance (i.e., in the same ring) from the last relay (or from the initial sender) to compute a forwarding delay in a same range of values. In the part of the network depicted in Figure 2, nodes N_1, N_2 and N_3 will thus compute a forwarding delay in a same range of values. This delay will be less than the one computed by N_4, which itself will be less than the one computed by N_5. Moreover, a node perceiving that all of its neighbors have already received the message it plans to forward will cancel its forwarding process, and will trigger it when it is notified of the arrival of a new node in its vicinity. Thus in our scenario, node N_5 will not forward advertisement A, because this advertisement is rebroadcast by node N_4 first. If we consider that all the nodes have the same communication range of radius R, we can deduce, based on geometric properties, that, in favorable conditions, only 3 nodes will forward advertisement A the first time [15]. Consequently at hop n, in favorable conditions the number

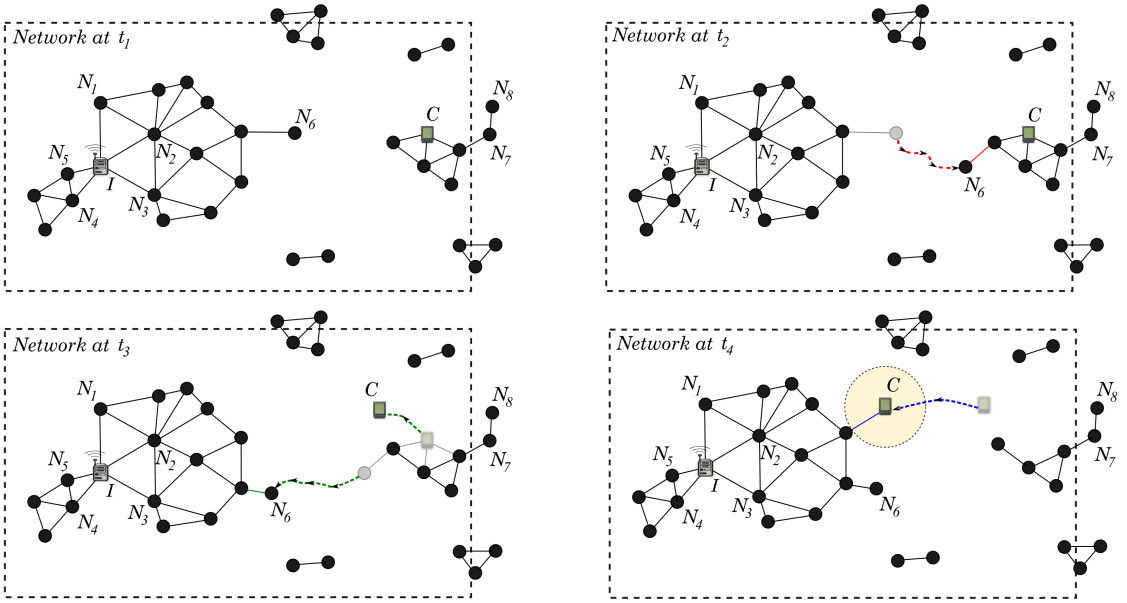


Figure 2. Opportunistic communication in a DMANET with OLFServ.

of forwarders will be $3 \times n$, and in the worst conditions (i.e., when the selected forwarders moved before forwarding their message, and become out of reach of each other), the number of forwarders will be $\sum_{i=0}^{n} 6^n$. This property is thus independent of the density of the network.

By implementing the "store, carry and forward" principle and by exploiting the nodes' mobility and contact opportunities, advertisement A will be propagated in the whole area specified by the infostation, and only in this area. Indeed, the self-pruning heuristics implemented in our protocol prevent mobile devices from forwarding messages outside the area specified in the headers of these ones. For instance, node N_6 that left the island of infostation I at time t_1 and joined that of client C at time t_2 will broadcast advertisement A in this new island. This message will be then broadcast by the other nodes of this island whether it is still valid (i.e., the number of hops is greater than zero and the lifetime has not expired yet), except by node N_7 because it is outside the area specified by infostation I. Thus, node N_8 will not receive message A.

b) Service invocation: After discovering the service offered by infostation I, client node C can invoke this service by sending an invocation request including namely the ID of the infostation, the location of this one, and its own location. Let us also consider that client C knows its speed and its direction and that it has also included them in the request it sent, thus allowing to compute with a better accuracy the area where it is expected to be when it will receive the response. Indeed, when the speed and the direction (or the destination) are unknown, the "expected area" is a circle whose center is the current position of the client and whose radius is proportional to a predefined speed (of about 2 m/s for pedestrians) and to the time expected for the response delivery (this time is estimated from the request delivery time). The notion of "expected area" was introduced in [16]. In contrast, when the speed and the direction are known, the "expected area" is a circle centered on the position computed from the speed and the direction indicated by the client, and whose radius is proportional to the inaccuracies of both the speed and the forwarding time (see the dotted circle in Figure 2).

The request sent by C will be received by intermediate nodes and broadcast by these ones towards infostation I following a forwarding scheme that is quite similar to the discovery forwarding scheme presented previously. The difference between these two schemes resides in the number of nodes that will rebroadcast the messages. Indeed, since the invocation process is usually achieved using a unicast communication scheme, we have introduced additional self-pruning heuristics in comparison to the service discovery process in order that only the nodes closer to the destination than the previous hop can forward the message towards the destination. Thus, the area where the message is forwarded is progressively refined until reaching the destination, and the number of messages that are replicated in the network is reduced while having a good message delivery ratio. A node, receiving a message from a neighbor node closer to the message's recipient than itself, will store the message locally and will forward this message later when it becomes closer to the recipient than this neighbor. For example N_7 and N_8 will not broadcast the request sent by node C at time t_2 because they are farther than C from infostation I. This invocation request will be received by node N_6 at time $t_2 + \Delta t$. If N_6 joins the island of infostation I at time t_3 as shown in Figure 2, it will broadcast this request in this island because it will discover new neighbors that have not received this message yet. These neighbors will then forward this request towards infostation I.

If client C has specified its location, its speed and its possible direction of movement, OLFServ can estimate the area where C is expected to be when it should receive the response from I. So when the response is returned, this area is specified in a header of this message. The response will be then

routed towards this "expected area" using a forwarding scheme comparable to that used for the invocation. When the message has reached the "expected area", it will be disseminated in this area following a broadcast scheme comparable to that used for service discovery. This technique is used since the position of the client cannot be computed with a good accuracy due to the delay induced by opportunistic communication. When a mobile device receives a response for an invocation it has previously stored locally, it stops forwarding this request in the network. In our scenario (Figure 2) the response will be routed towards node C by nodes N_2, N_3 or N_1 because they are closer to the "expected area" than I. Moreover, if an invocation request reaches the provider within a short amount of time (i.e., if a end-to-end route is very likely to exist between the client and the provider), OLFServ tries to follow the same route by applying source routing. If the source routing process failed because an intermediate node has moved, then the node perform a broadcast towards the destination as mentioned before. Finally, if a node stored previously a response for the request sent by client C, it will send back this response (if it is still valid) instead of forwarding the request towards infostation I. For instance, N_2 can return to client C the copy of the response it holds locally, instead of forwarding the request to I. Thus, the number of message roaming in the network is reduced and the service invocation responsiveness is improved. The same process is applied when a client is looking for a service: an intermediate node can send back to the client the advertisement it holds locally that "matches" the service discovery request sent by the client.

Algorithm 1 Reaction on message reception.

Input:
m: the incoming message
t: the current time
C, Δ, K_m, N

if ($m \in \Delta$ & $N \subseteq K_m$) **then**
  $\Delta \leftarrow \Delta - \{m\}$
else
  if ($\exists\ p \in C$ / p is response for m & $p[lifetime] > t - p[date]$ & $p[hops] > 0$) **then**
    compute forwarding delay for p
    $\Delta \leftarrow \Delta \cup \{p\}$
  else
    if ($\exists\ k \in C$ / m is response for k) **then**
      $C \leftarrow C - \{k\}$
      if ($k \in \Delta$) **then**
        $\Delta \leftarrow \Delta - \{k\}$
        if ($k \in Q_s$) **then**
          $Q_s \leftarrow Q_s - \{k\}$
        else
          $Q_b \leftarrow Q_b - \{k\}$
        end if
      end if
      if ($t - k[reception\_date] < \varepsilon$) **then**
        $m[source\_routing] \leftarrow k[L_{relay}]$
      end if
    end if
    if ($m[lifetime] > t - m[date]$ & $m[hops] > 0$) **then**
      $C \leftarrow C \cup \{m\}$
      $m[reception\_date] \leftarrow t$
      $K_m \leftarrow K_m \cup \{m[K_m]\}$
      if ($N \not\subseteq K_m$) **then**
        compute forwarding delay for m
        $\Delta \leftarrow \Delta \cup \{m\}$
      end if
    end if
  end if
end if

C. Specification of the protocol

The remainder of this section presents how OLFServ reacts when one of the above-mentioned events occurs.

1) Notations: The location of a node is subsequently identified as L, the one of the last relay as L_{relay} and the one of the destination as $L_{recipient}$. The one-hop neighborhood of a node is referred to as N. The local cache of a node is identified as C. Q_s and Q_b are outgoing queues for the messages that must be sent using source routing techniques and for the messages that must be broadcast respectively. K_m refers to the set of nodes that are known to have received message m. Δ is the set of messages that must be forwarded and for which a forwarding delay has been computed. Finally, the messages headers can include several properties (the location of the recipient, the location of the sender, a date of emission, a lifetime, a maximum allowed number of hops, the geographical area where the message can be disseminated, etc.). A given property of a message m is identified as $m[property]$.

2) Message reception: When receiving a message m, Algorithm 1 is applied. First, if a node receives from one of its neighbors a message it plans to forward, it checks if all of its neighbors have received this message. If so, it cancels its forwarding process. If the node has in its cache an advertisement p for the service discovery request m (or a response p for the invocation request m) then the node is expected to forward p if this one is still valid. A forwarding delay is computed for message p, and p is put in the set of messages that must be sent. Otherwise, if m is a response for an invocation request k (or if m is an advertisement for a discovery request k), k is removed from the local cache in order not to be forwarded later, as well as from the set of messages that must be forwarded. If message m is still valid and if the number of hops is greater than 0, message m is put in the local cache, and the set K_m is updated (i.e., the set of nodes that are known to have received message m yet). Message m is put in the set of messages that must be forwarded and a forwarding delay is computed for m. When the forwarding delay δ_m expires, Algorithm 3 will be applied.

3) Computation and expiration of the forwarding delay: Each mobile device computes a forwarding delay for each message it receives. This delay prevents close devices from forwarding messages simultaneously. As mentionned before, in OLFServ the forwarding delay has both a random component and a component that is inversely proportional to the distance from the previous relay. So as to compute this forwarding delay, the wireless communication range of each device has been divided in several rings (see Figure 2), so that the delays computed by hosts in ring i are greater than those computed by hosts in ring $i+1$. The mobile hosts of a given ring are considered as equivalent regarding the spatial propagation of messages. The algorithm used to compute the forwarding delay is described in Algorithm 2. This algorithm has mainly three parameters: the wireless communication range (W), the ring size (rs) and α. This last parameter has

Algorithm 2 Computation of the forwarding delay.

Input:
m: the incoming message
rs: the ring size
R: the ring number
δ: the default forwarding period
W: the wireless communication range
Output: δ_m: the forwarding delay for message m
1: $R \leftarrow floor((W - distance(L, m[L_{relay}]))/rs)$
2: $\delta_m \leftarrow min(\delta, \alpha * random(R * rs, (R+1) * rs))$

Algorithm 3 Expiration of the forwarding delay.

Input:
t: the current time
m: the message
C, N, K_m, Q_b, Q_s
1: **if** $(N - K_m \neq \emptyset$ & *in* $m[area]$ & $m[lifetime] > t - m[date]$ & $m[hops] > 0)$ **then**
2: **if** $(m[recipient] \neq "*")$ **then**
3: $d_{this \rightarrow recipient} \leftarrow distance(L, m[L_{recipient}])$
4: $d_{relay \rightarrow recipient} \leftarrow distance(m[L_{relay}], m[L_{recipient}])$
5: **if** $(d_{this \rightarrow recipient} \leq d_{relay \rightarrow recipient})$ **then**
6: $m[area] \leftarrow (m[L_{recipient}], d_{this \rightarrow recipient})$
7: $m[K_m] \leftarrow m[K_m] \cup K_m$
8: $m[L_{relay}] \leftarrow L$
9: $m[nb\ hops] \leftarrow m[nb\ hops] - 1$
10: **if** $(t - m[date] < \varepsilon)$ **then**
11: $Q_s \leftarrow Q_s \cup \{m\}$
12: $\Delta \leftarrow \Delta - \{m\}$
13: **else**
14: $Q_b \leftarrow Q_b \cup \{m\}$
15: $\Delta \leftarrow \Delta - \{m\}$
16: **end if**
17: **end if**
18: **else**
19: $m[K_m] \leftarrow m[K_m] \cup K_m$
20: $m[L_{relay}] \leftarrow L$
21: $m[nb\ hops] \leftarrow m[nb\ hops] - 1$
22: $Q_b \leftarrow Q_b \cup \{m\}$
23: $\Delta \leftarrow \Delta - \{m\}$
24: **end if**
25: **end if**

been introduced in order to define a relevant delay δ_m: the delay in the largest ring is of the order of a few milliseconds, while in the smallest ring it is of the order of a few seconds typically.

When the forwarding delay of a given message has expired, Algorithm 3 is applied. If there are new nodes in the one-hop neighborhood, if the client is in the area where the message can be disseminated, if the message is still valid and if the message has next hops, the message is then considered as being forwardable. The headers of the message are then updated. If the destination is known, the area where the message can be propagated is updated in order to refine this area progressively until reaching the destination. Moreover if the destination is known, the mobile device checks whether it is closed to the destination than the last forwarder, and if so, it updates the number of hops, the location of the last forwarder with its own location and the set of nodes that have already received the message, and puts the message in the outgoing message queue. If the message has expired or if the number of hops equals to 0, the message is removed from the local cache.

4) Location changes: When reaching a given location, a mobile host can trigger the forwarding of some messages. For

Algorithm 4 Location changes.

Input:
t: the current time
m: the message that must be forwarded
C,K_m
1: **if** $(m[lifetime] > t - m[date]$ & $m[hops] > 0$ & $N \not\subseteq K_m)$ **then**
2: **if** $(m[type] = response$ & L *in* $m[expected\ area])$ **then**
3: $m[recipient] \leftarrow "*"$
4: **end if**
5: compute forwarding delay for m
6: **end if**

Algorithm 5 Detection of new neighbor nodes.

Input:
t: the current time
n: the new neighbor
C, N
1: $N \leftarrow N \cup \{n\}$
2: **for all** $m \in C$ **do**
3: **if** $(m[lifetime] > t - m[date]$ & $m[hops] > 0)$ **then**
4: **if** $(n \notin K_m$ & *in* $m[area])$ **then**
5: compute forwarding delay (m)
6: **end if**
7: **else**
8: $C \leftarrow C - \{m\}$
9: **end if**
10: **end for**

instance, a mobile host that was far from the recipient of a message it received can trigger the emission of this message when it is at a given distance from the recipient. Similarly, when entering the area where a client is likely to be receiving its service response, a mobile host, acting as an intermediate node, can both update the message headers in order that this message can be broadcast in this whole area and trigger its emission. When the mobile host has reached a given location, Algorithm 4 is executed. We change the status of the response in order that it is broadcast by the node in the whole area specified by the provider. And for each message when we become closer to the destination than the previous node (the node from which we have received the message), we trigger a message emission.

5) New neighbor detection: When a new neighbor node is discovered, the mobile host computes a forwarding delay for all the messages that are still valid, that have next hops, if the new neighbor is not in the the list of nodes that have already received the message and if the mobile host is in the area where the message can be propagated. A new forwarding delay is computed in order to prevent the emission of the same messages by different nodes that simultaneously discover the new neighbor node in their one-hop neighborhood.

IV. Experiments and results

In order to evaluate our protocol, we conducted a series of simulations using the Madhoc simulator (http://www-sop.inria.fr/members/Luc.Hogie/madhoc), a metropolitan ad hoc network simulator that features the components required for both realistic and large-scale simulations, as well as the tools essential to an effective monitoring of the simulated applications. This simulator, which is written in Java, allows us to run our middleware platform on it. In the current scenarios

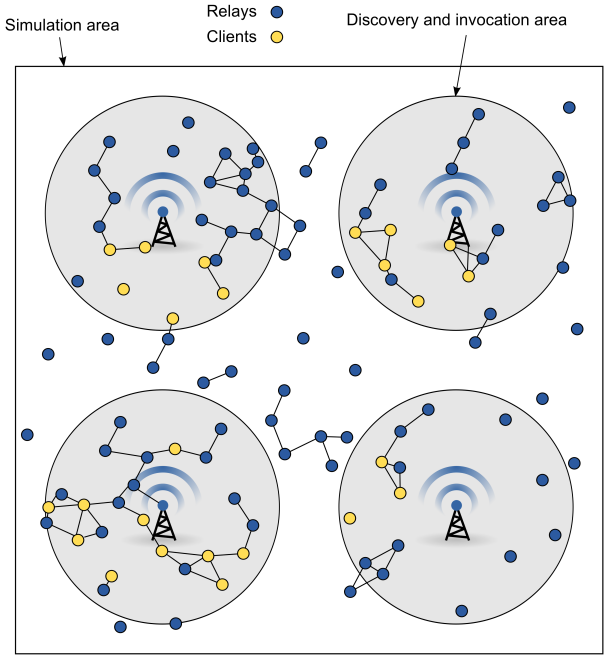


Figure 3. Simulation environment.

we focus on, service providers are fixed infostations deployed in a city, while clients are devices carried by humans.

A. Experiments and simulation setup

The simulation environment we consider is depicted in Figure 3. It is an open area of about 1 km^2. Four infostations offering two different services are deployed in this environment. These services can be discovered and invoked in a circular area of a radius of 200 m. The first service delivers the day's weather forecast, while the second provides an access to a "yellow page" service, which can be invoked by nomadic people in order to find restaurants, shops, etc. Mobile clients are thus expected to submit the same request to the first service and different ones to the second service. In our simulations, we have considered successively 50, 100, 500 and 1000 pedestrians carrying a PDA (Personal Digital Assistant) equipped with both a Wi-Fi interface and a GPS receiver. The communication range of both mobile devices and infostations varies from 60 to 80 m. Some of the pedestrians move randomly, while others follow predefined paths. Each pedestrian moves at a speed between 0.5 and 2 m/s. In our simulations, 30% of the mobile devices act as clients of the above-mentioned services, whereas the others only act as intermediate nodes. The service providers are expected to broadcast service advertisements every 30 seconds when mobile devices are in their vicinity. After discovering the services they are looking for, the clients invoke these services every 3 minutes. In our experiments, we have assigned to all the messages a lifetime of 5 minutes and a maximum number of hops of 8. We present below the results we obtained for OLFServ in these various configurations, and we compare OLFServ with the Epidemic Routing Protocol (EPR) defined by Vahdat and Becker [17]. The objective of these experiments was to measure the ability to satisfy the client service discovery and invocation efficiently with a small number of message copies.

B. Results

Figure 4 shows the service discovery delays we have observed in the various simulation setups we have considered. As expected, we can see that the discovery delays decrease when the number of nodes increases. Indeed, in a dense environment the connectivity disruptions are less frequent, and the impact of the opportunistic communications are reduced. The discovery process can be perceived as a long process. For instance, only 70% of the clients have discovered the service they require after 20 minutes in the second setup (30 clients and 70 relays). However, it should not be forgotten that the services can be discovered and invoked by the clients only in restricted areas and not in the whole environment (see Figure 3), with the consequence that several minutes may elapse before the clients have reached the restricted area of the service they are looking for. However, the speed of discovery inside this restricted area is significantly greater: we have observed that, in most of the situations, the discovery time is less than 10 seconds after the client has entered the area of the service it requires, and that it lasts about 1 minute in the worst case. Finally, the services are discovered more quickly with OLFServ than with the epidemic routing protocol. In OLFServ the service advertisements are broadcast by the mobile nodes, whereas in the EPR, the nodes must first exchange summary vectors with each of their neighbors before forwarding the service advertisements themselves, thus introducing a latency in the discovery process.

Figure 5 and Figure 6 present the simulation results for the two kinds of services considered (the "weather forecast" service S1 and the "yellow pages" service S2). Figure 5 gives the average number of emissions for a service advertisement (for S1 and S2) with OLFServ and with EPR. One can observe that the number of emissions increases drastically with EPR, while it remains relatively constant with OLFServ. Indeed, in EPR, when two hosts come into communication range of one another, they exchange their summary vectors to determine which messages stored remotely have not been seen by the local host. In turn, each host then requests copies of messages that it has not seen yet. In contrast in OLFServ, service advertisements are broadcast and not sent using a unicast communication model. Moreover, only a subset of the neighbor nodes are expected to rebroadcast these advertisements in turn. For S2, the number of emissions of a given service invocation request is less than the half of the number of emissions of service advertisements (see Figure 6). These results are consistent with those expected. Indeed, the invocation requests are broadcast only by the nodes closer to the destination at each hop. It must be noticed that the number of emissions of invocation requests for S1 is less than that for S2. Again, the results are consistent with those expected: all the clients interested in the "weather forecast" service submit the same request, and obtain in return the same response during the simulation. The mobile nodes that have stored a request and the associated response are able to establish a correlation

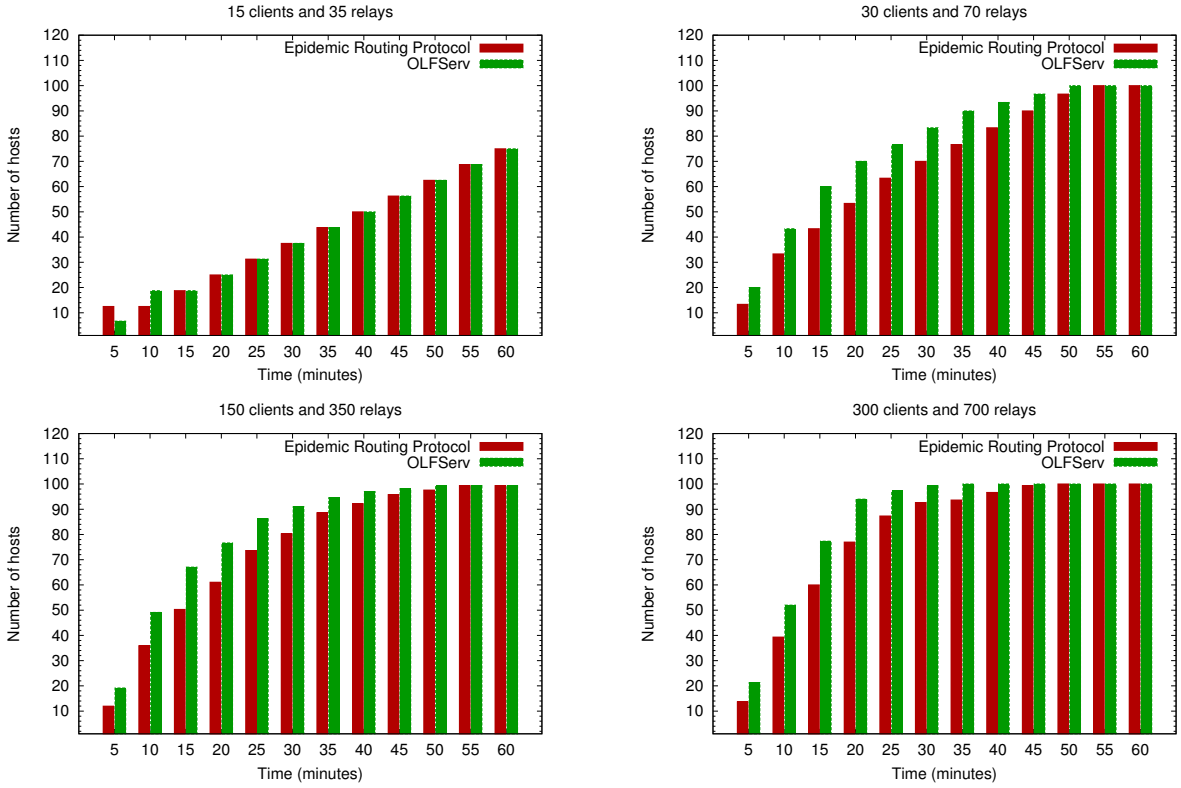


Figure 4. Service discovery delays.

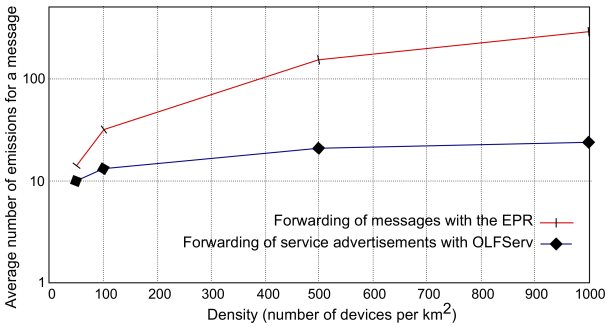


Figure 5. Service advertisement with OLFServ and EPR.

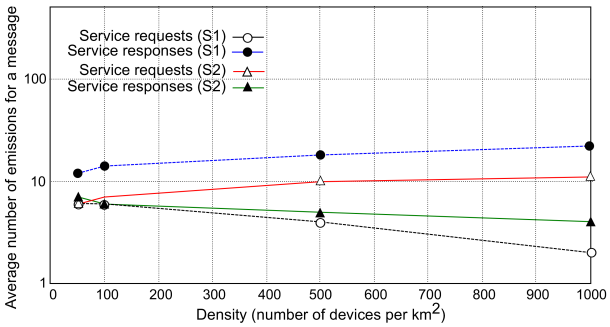


Figure 6. Service invocation with OLFServ.

between these messages, and are expected to send back to the client the stored response when they receive a new similar request. The number of requests for S1 decreases according to the number of clients. Such a phenomenon can be explained by the fact that a request is not forwarded by a node towards the destination if this node has already obtained the response associated with this request. This correlation techniques is further detailed in [5]. Finally, it must be noticed that the mobility of nodes between the successive invocations does not allow benefiting from source routing when forwarding a request towards a provider. Nevertheless, source routing has proved its efficiency in the forwarding of the responses, as shown in Figure 5. Thus, the number of messages sent in the network is reduced while offering a better service provision (see Table I).

As shown in Table I, the number of clients that have discovered the service they are looking for is greater with EPR than with OLFServ. Nevertheless, the invocation success ratio with EPR is less than with OLFServ. Indeed, with OLFServ messages are routed only in the areas where the services can be discovered and invoked, whereas with EPR, messages are routed in the whole simulation area. Consequently, with EPR, services can sometimes be discovered by the clients, but not invoked successfully due to the mobility of intermediate nodes, to the periodic exchange of messages (every 20 seconds) and to the fixed number of hops. In contrast, with OLFServ, messages are forwarded few milliseconds after their reception instead of being forwarded periodically. OLFServ thus offers a good responsiveness and delivery ratio while producing a lower network load.

	EPR(50)	EPR(100)	EPR(500)	EPR(1000)	OLFServ(50)	OLFServ(100)	OLFServ(500)	OLFServ(1000)
Average delay of successful invocations to service S1 (seconds)	120	100	60	40	1,02	0,58	0,43	0,42
Average delay of successful invocations to service S2 (seconds)	120	100	60	40	3,32	2,84	2,43	2,42
Average ratio of successful invocations	0.78	0,84	0,92	0,96	1	1	1	1

Table I
SIMULATION RESULTS FOR SERVICE INVOCATION.

V. RELATED WORK

Our work on OLFServ is related to works on broadcast protocols [18], [19]. Indeed, some techniques that aim at reducing the number of message forwarders are adapted or integrated to the specific context of service provision in opportunistic networks.

However, the research works that follow the same objectives as OLFServ are mainly led in the opportunistic networking and/or delay/disrupted tolerant networking domain. One of the first protocol in this domain is the Epidemic Routing Protocol [17], which can in a way be assimilated to a simple flooding, not suitable for environments with high density regions, since it would generate too much network traffic and could even lead to network congestion. This drawback is addressed by protocols implementing methods aiming to assess the capability of a neighbor node to contribute to the delivery of a given message. These methods usually use a probabilistic metric, often called delivery predictability, that reflects how a neighbor node will be able to deliver a message to its final recipient [20]. Before forwarding (or sending) a message, a mobile host asks its neighbors to infer their own delivery probability for the considered message, and then compares the probabilities returned by its neighbors and chooses the best next carrier(s) among them. In CAR [12] and GeOpps [8], the delivery probabilities are computed using both utility functions and Kalman filter prediction techniques. CAR assumes an underlying MANET routing protocol that connects together nodes in the same MANET cloud. To reach nodes outside the cloud, a sender looks for the node in its current cloud with the highest probability of delivering the message successfully to the destination. GeOpps, which is a geographical delay-tolerant routing algorithm, exploits the pieces of information provided by the vehicles' navigation system in order to route the messages to a specific location. Like CAR, HiBOp [13] also exploits context information in order to compute delivery probabilities. However, HiBOp can be perceived as being more general than CAR since it does not require an underlying routing protocol, and because it is also able to exploit context for those destinations that nodes do not know. HiBOp exploits history information in order to improve the delivery probability accuracy, and does not make predictions as CAR. Propicman [9], as for it, also exploits context information and uses the probability of nodes to meet the destination, and infers from it the delivery probability, but in a different way. When a node wants to send a message to another node, it sends to its neighbor nodes the information it knows about the destination. Based on this information, the neighbor nodes compute their delivery probability and return it. In Prophet [10], the selection of the best neighbor node is based on how frequently a node encounters another. When two nodes meet, they exchange their summary vectors, which contain their delivery predictability information. If two nodes do not meet for a while, the delivery predictability decreaces. When a node wants to send a message to another node, it will look for the neighbor node that has the highest amount of time encountering the destination, meaning that has the highest delivery predictability to the destination. Furthermore, this property is transitive. Unlike OLFServ, most of the above-mentioned protocols rely on an history of contacts and a prediction of encounters in order to select the best next carrier(s). Computing such an history and a prediction is a tricky problem, especially in environments composed of numerous mobile devices that move following irregular patterns, such as those hold by pedestrians in a city. Although they implement various strategies aiming to select the next best carriers(s) to deliver a given message, the above-mentioned protocols are not suited to service discovery. Indeed, they implement neither self-pruning heuristics making it possible for mobile nodes to decide if they should rebroadcast a message according to their neighborhood perception, nor methods allowing to designate which subset of neighbor nodes must rebroadcast a message. If used to broadcast service advertisements or service discovery requests network-wide, they will probably induce a storm of messages and perhaps a network congestion.

Geographic routing protocols, such as GeRaf [21], LAR [16] and Dream [22], propose forwarding techniques similar to those implemented in OLFServ. Once a node has a message to send, it broadcasts it while specifying its own location and the location of the destination. All the nodes in the coverage area will receive this message and will assess their own capability to act as a relay, based on how close they are to the destination. Dream and LAR also propose some solutions in order to improve the message delivery in MANETs. For instance, based on location information, they can compute the area where the mobile clients are expected to be when they receive their messages. Nevertheless, on contrary to OLFServ, these protocols do not implement the "store, carry and forward" principle and therefore are not suitable for disconnected MANETs.

VI. CONCLUSION

The vision of opportunistic computing is to provide mobile users with pervasive access to software services without rely-

ing on a fixed infrastructure but rather exploiting direct radio contacts between mobiles devices in a disconnected MANET. Opportunistic transmissions are performed during these contacts, enabling routing of messages between services clients and service providers. In this context, the work described in this paper focused on routing in the case where service providers are fixed infostations and where devices in the network are endowed with the capacity to geolocalize themselves. We proposed a new forwarding protocol called OLFServ, suited for service provision in disconnected MANETs. This protocol implements several self-pruning heuristics aiming to efficiently control the dissemination of service advertisements and service discovery requests, as well as to perform a geographic and source-based routing allowing cost effective delivery of service invocation requests and responses. Simulation results show that OLFServ outperforms epidemic routing in networks composed of numerous mobile devices moving randomly with respect to delivery delay, delivery ratio and number of emissions (reflecting the network throughput). In the future, we would like to investigate new complementary techniques, such as geometric localized forwarding and spanning trees, in order to forward a message from a source to a destination along different paths while reducing again the delay and the message copies, especially when some partitions of the network are temporarily stable.

REFERENCES

[1] N. Le Sommer and Y. Mahéo, "OLFServ: an Opportunistic and Location-Aware Forwarding Protocol for Service Delivery in Disconnected MANETs," in *Proceedings of the 5th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM'2011)*, Lisbon, Portugal, pp. 115–122, Xpert Publishing Services, Nov. 2011.

[2] H. A. Nguyen and S. Giordano, "Routing in Opportunistic Networks," *International Journal of Ambient Computing and Intelligence*, vol. 1, no. 3, pp. 19–38, 2009.

[3] M. Conti, S. Giordano, M. May, and A. Passarella, "From Opportunistic Networks to Opportunistic Computing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 126–139, 2010.

[4] Y. Mahéo, N. Le Sommer, P. Launay, F. Guidec, and M. Dragone, "Beyond Opportunistic Networking Protocols: a Disruption-Tolerant Application Suite for Disconnected MANETs," in *Proceedings of the 4th Extreme Conference on Communication (ExtremeCom 2012)*, Zürich, Switzerland, ACM, Mar. 2012.

[5] S. Ben Sassi and N. Le Sommer, "Towards an Opportunistic and Location-Aware Service Provision in Disconnected Mobile Ad Hoc Networks," in *Proceedings of the International Conference on Mobile Wireless Middleware, Operating Systems, and Applications (Mobilware 2009)*, Berlin, Germany, vol. 7 of *LNICST*, pp. 396–406, Springer-Verlag, Apr. 2009.

[6] S. Y. Ni, Y. C. Tseng, Y. S. Chen, and J. P. Sheu, "The Broadcast Storm Problem in a Mobile Ad Hoc Network," in *Proceedings of the 5th International Conference on Mobile Computing and Networking (MobiCom 99)*, Seattle, Washington, USA, pp. 151–162, ACM/IEEE CS, Aug. 1999.

[7] Y. Mahéo and R. Said, "Service Invocation over Content-Based Communication in Disconnected Mobile Ad Hoc Networks," in *Proceedings of the International Conference on Advanced Information Networking and Applications (AINA 2010)*, Perth, Australia, pp. 503–510, IEEE CS, Apr. 2010.

[8] I. Leontiadis and C. Mascolo, "GeOpps: Geographical Opportunistic Routing for Vehicular Networks," in *Proceedings of the International Symposium on a World of Wireless, Mobile and Multimedia Networks – AOC Workshop (WoWMoM 2007)*, Helsinki, Finland, IEEE CS, June 2007.

[9] H. A. Nguyen, S. Giordano, and A. Puiatti, "Probabilistic Routing Protocol for Intermittently Connected Mobile Ad hoc Network (PROPICMAN)," in *Proceedings of the International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007)*, Helsinki, Finland, IEEE CS, June 2007.

[10] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic Routing in Intermittently Connected Networks," in *Proceedings of the International Workshop on Service Assurance with Partial and Intermittent Resources (SAPIR 2004)*, Fortaleza, Brazil, vol. 3126 of *LNCS*, pp. 239–254, Springer Verlag, Apr. 2004.

[11] F. Guidec and Y. Mahéo, "Opportunistic Content-Based Dissemination in Disconnected Mobile Ad Hoc Networks," in *Proceedings of the International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2007)*, Papeete, French Polynesia, pp. 49–54, IEEE CS, Nov. 2007.

[12] M. Musolesi and C. Mascolo, "CAR: Context-Aware Adaptive Routing for Delay Tolerant Mobile Networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 2, pp. 246–260, 2009.

[13] C. Boldrini, M. Conti, I. Iacopini, and A. Passarella, "HiBOp: a History-Based Routing Protocol for Opportunistic Networks," in *Proceedings of the International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007)*, Helsinki, Finland, pp. 1–12, IEEE CS, June 2007.

[14] L. Hogie, P. Bouvry, F. Guinand, G. Danoy, and E. Alba, "A Bandwidth-Efficient Broadcasting Protocol for Mobile Multi-hop Ad hoc Networks," in *Proceedings of the 5th International Conference on Networking (ICN 2006)*, Mauritius, IEEE CS, Apr. 2006.

[15] X. Liu, X. Jia, H. Liu, and L. Feng, "A Location-Aided Flooding Protocol for Wireless Ad Hoc Networks," in *Proceedings of the International Conference on Mobile Ad-Hoc and Sensor Networks (MSN 2007)*, Beijing, China, vol. 4864 of *LNCS*, pp. 302–313, Springer-Verlag, Dec. 2007.

[16] Y.-B. Ko and N. H. Vaidya, "Location-Aided Routing (LAR) in Mobile Ad Hoc Networks," *Wireless Networks*, vol. 6, pp. 307–321, 2000.

[17] A. Vahdat and D. Becker, "Epidemic Routing for Partially Connected Ad Hoc Networks," tech. rep., Duke University, 2000.

[18] B. Williams and T. Camp, "Comparison of Broadcasting Techniques for Mobile Ad Hoc Networks," in *Proceedings of the 3rd International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2002)*, Lausanne, Switzerland, pp. 194–205, ACM, June 2002.

[19] I. Stojmenovic and J. Wu, *Mobile Ad Hoc Networking*, ch. 7: Broadcasting and Activity-Scheduling in Ad Hoc Networks, pp. 205–229. Wiley, 2004.

[20] J. Wu and F. Dai, "Broadcasting in Ad Hoc Networks Based on Self-Pruning," in *Proceedings of the Joint Conference of the IEEE Computer and Communications Societies (InfoComm 2003)*, San Francisco, California, USA, vol. 3, pp. 2240–2250, IEEE CS, 2003.

[21] M. Zorzi and R. R. Rao, "Geographic Random Forwarding (GeRaF) for Ad Hoc and Sensor Networks: Multihop Performance," *IEEE Transactions on Mobile Computing*, vol. 2, pp. 337–348, 2003.

[22] S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, "A Distance Routing Effect Algorithm for Mobility (DREAM)," in *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom'98)*, Dallas, Texas, USA, pp. 76–84, ACM, Oct. 1998.

Association Control for Wireless LANs: Pursuing Throughput Maximization and Energy Efficiency

Oyunchimeg Shagdar*, Suhua Tang†, Akio Hasegawa†, Tatsuo Shibata†, Masayoshi Ohashi†, and Sadao Obana‡
Email: oyunchimeg.shagdar@inria.fr, {shtang, ahase, shibata, ohashi}@atr.jp, and obana@cs.uec.ac.jp
* IMARA Project-Team, INRIA Rocquencourt
Domaine de Voluceau, B.P. 105 78153, Le Chesnay, FRANCE
† ATR Adaptive Communications Research Laboratories
Hikaridai, Keihanna Science City, Kyoto, JAPAN
‡ Graduate School of Informatics and Engineering, University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, JAPAN

***Abstract*—Because the access points (APs) and the stations (STAs) of a community access network are deployed at the users' desired places, the APs and STAs tend to concentrate in certain areas. A concentration of STAs often results in the AP(s) and STAs in that particular area suffering from severe congestion. A concentration of APs, on the other hand, may cause energy wastage. While a number of association control schemes are proposed to alleviate congestion in WLANs, the existing schemes do not necessarily maximize throughput and do not consider energy consumption. In this paper, we analytically formulate the network throughput as the multiplication of the success probability, frame transmission rate, and channel air-time ratio. The second and third components can easily be monitored and controlled based on measurements of local link and channel condition using the off-the-shelf WLAN devices. On the other hand, the first component, success probability is a function of the number of contending nodes that is extremely difficult to monitor in overlapping WLANs. Due to this reason, we extend our theoretical study and show that success probability can be indirectly maximized by controlling air-time ratio. Finally, we propose an association control scheme that aims at maximizing throughput and reducing energy consumption by taking account of the multiplication of frame transmission rate and air-time ratio. The proposed scheme is evaluated by computer simulations and testbed experiments conducted under real-world complex scenarios with UDP and TCP traffic. Both the simulations and actual implementations confirm the correctness of the theoretical work and the effectiveness of the proposed scheme.**

Index Terms*—*association control*; *throughput maximization*; *success probability*; *air-time ratio*; *congestion alleviation*; *energy efficiency

I. Introduction

Due to the increasing popularity of WLAN technology, users (i.e., STAs) are often in the vicinity of one or more APs deployed at offices, school campuses, hotspot areas (e.g., cafes, train stations, airports), and individuals' homes. Such a widespread deployment of WLAN triggered a launch of community access networks, including FON [2], which are built exploiting user-owned APs. As community access networks enable users to enjoy ubiquitous Internet access, it has the potential to play an important role in the future networking paradigm.

The APs of a community access network are deployed at the users' desired places. Therefore, the fundamental difference

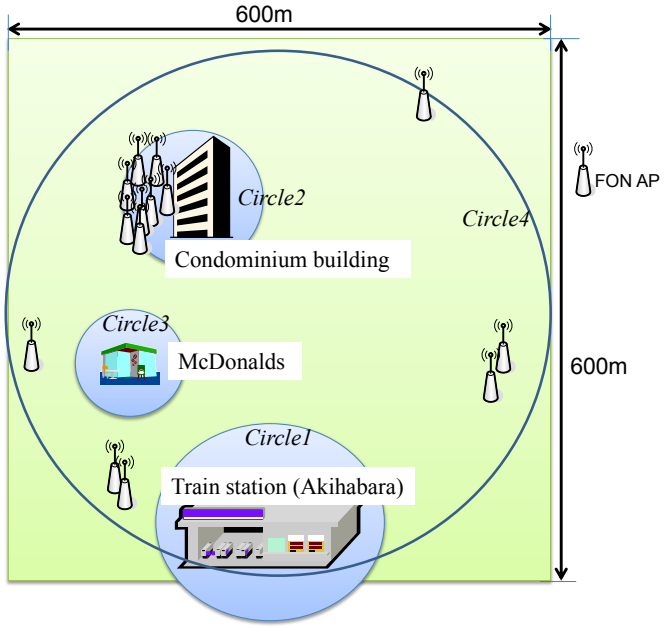


Figure 1. A map showing locations of FON APs in 600m×600m area in Tokyo. 8 out of the total 16 APs are concentrated in a small area (1/8 of the overall area). (The information is taken from maps.fon.com on July 20, 2010)

of community networks from, e.g., enterprise wireless access networks is that a systematic placement of the APs is not possible. In community networks, APs are often concentrated in areas such as a residential street, and their distribution is highly non-uniform. Fig. 1 shows a FON map for an approximately 600m×600m area in Tokyo (near Akihabara station). Out of the total 14 APs installed in the overall area, 8 APs are concentrated in a small area in the upper left part of the map (approximately 1/8 of the overall area). The majority of these APs are in fact deployed in a condominium building, where residential homes, a café, conference spaces, and a fitness centre are. The remaining 6 APs are distributed in the rest 7/8 of the overall area as shown in the figure. STAs (i.e., users) are, on the other hand, expected to be concentrated in public places, such as cafés and train stations. Therefore, it is clear that STAs and APs are not necessarily concentrated

in a same area. A concentration of STAs often results in the AP(s) and STAs in that particular area suffering from severe congestion [3]-[4]. A concentration of APs, on the other hand, may cause energy wastage [5].

Congestion can be effectively alleviated by a proper association control. Indeed, a number of association control schemes are proposed for WLANs, mainly aiming at load balancing among cells under different definitions of load (e.g., load is defined as the number of nodes in [3], [6], channel condition in [7], and traffic rate in [8], [4]).

In [1], we analytically formulated the throughput maximization problem for overlapping WLANs, and showed that the existing schemes do not necessarily operate towards throughput maximization. The throughput can be expressed as the multiplication of three components: success probability, frame transmission rate, and air-time ratio (ATR). The success probability is the probability of collision-free transmission and it is a function of the number of contending nodes. The frame transmission rate is the speed of the data transmission, and it is a function of the link quality. ATR is the ratio of the channel busy time to the total time. While link quality and ATR can easily be monitored in WLAN systems, counting the number of contending nodes is extremely difficult because the nodes can be beyond the transmission range of each other. Thus in [1], we proposed an association control scheme that aims at throughput maximization by maximizing the multiplication of frame transmission rate and ATR. Success probability is indirectly maximized by controlling ATR. Moreover, because the existing association control schemes aim at load balancing among cells, they tend to use all the existing APs. However, especially in AP-concentrated areas, throughput maximization can be achieved by utilizing fewer APs. Such an association control provides a positive impact on energy efficiency since the unused APs can be in the power-saving mode. Indeed the proposed scheme does not aim at load balancing and it can reduce the number of active APs without hampering the network throughput.

This paper extends our work by a detailed theoretical study and extensive performance investigations. More specifically, the contributions of this paper are:

- An analytical study that explains why and how the success probability can be maximized by controlling air-time ratio.
- An extensive investigation of the proposed scheme through testbed evaluations targeting UDP and TCP traffic in complex real-world scenarios.

The remainder of the paper is organized as follows. Section II formulates the throughput maximization problem and introduces the related works. In section III, we provide a theoretical study on ATR and its relation with success probability. Section IV introduces the proposed association control scheme that aims at maximizing network throughput and reducing energy consumption. Sections V and VI evaluate the proposed scheme through computer simulations and testbed experiments. Finally, we conclude the paper in section VII.

II. Problem Formulation and Related Work

The throughput of a node in a WLAN operating under DCF (Distributed Coordination Function) is expressed as follows [9], [10].

$$s = \frac{P_s P_{tr} E[P]}{(1 - P_{tr})\sigma + P_{tr} T} \quad (1)$$

Here, E[P] is the average data size, P_{tr} is the probability that there is at least one node transmitting in the sensing range of the node, and P_s is the probability of a successful transmission. σ is the slot time and T is the average time required for transmission of a data (includes DIFS and etc.). The numerator of (1) corresponds to the successfully transmitted payload length and the denominator is the total time. The former and the latter components of the denominator are the time that channel is idle and busy (either due to successful or unsuccessful transmissions), respectively.

The transmission of a data is successful if the frame is not collided and the frame does not contain errors (due to poor link quality). Thus, letting P_{sc} and P_{se} represent the probabilities of the former and the latter events, respectively, $P_s = P_{sc} \times P_{se}$. P_{sc} (in what follows we call P_{sc} as success probability) is expressed as

$$P_{sc} = \frac{\tau(1-\tau)^{n-1}}{1-(1-\tau)^n} \quad , \quad (2)$$

where n is the number of nodes. τ is the channel access probability, which is determined by the contention window size (CW) and the probability that a node has a frame to transmit [10].

Letting $r \equiv P_{se} \times E[P]/T$, we rewrite (1) as

$$s = P_{sc} \times r \times \frac{P_{tr} T}{(1 - P_{tr})\sigma + P_{tr} T} \quad . \quad (3)$$

$E[P]/T$, the average frame transmission rate, is variable if rate adaptation is deployed at MAC (Medium Access Control protocol), and it is fixed otherwise. If rate adaptation is deployed, the better the link quality (stronger RSSI), the higher is the selected transmission rate. Furthermore, if the rate adaption operates such that packet error rate (PER) is minimized (i.e., P_{se} is maximized), the second component of (3), r, depends mainly on the selected transmission rate, i.e., $r \approx E[P]/T$. The last component of (3) is the ratio of time the channel is determined to be busy to the total time. Letting ATR (air-time ratio) represent the last component, the throughput of a node is the multiplication of P_{sc}, r, and ATR.

Finally, the throughput maximization problem for a network that consists of multiple overlapping WLANs is the maximization of

$$S = \sum_{i=1}^{N} \sum_{j=1}^{n_i} s_{i,j} \quad , \quad (4)$$

where N is the number of cells, n_i is the number of STAs at the i^{th} cell, and $s_{i,j}$ is the throughput of the j^{th} STA at the i^{th} cell. It should be noted that the nodes that operate under the

same frequency channel and that are in each others' sensing range share the same P_{sc} and ATR regardless if they associated with a same or different APs.

In the traditional AP selection policy, a STA associates with the AP corresponding to the strongest RSSI. Thus, such a scheme takes account of only the second component of (3), r. However, increasing r alone does not necessarily increase the total throughput, especially when STAs distribution is highly non-uniform. In such a scenario, it is possible that most of the STAs select a same AP (because they are closer to that AP). As it can be seen in (2), P_{sc} decreases sharply with the increase of n (because the numerator approaches zero and denominator approaches 1). Thus, in such a case, the throughput of the traditional scheme is poor because of a small P_{sc} for the crowded cell(s) and likely a small ATR for the scarce neighboring cell(s). This suggests distributing STAs to the existing cells. Indeed several schemes are proposed to distribute the number of STAs among cells, and some of them take account of the link quality (RSSI in [3], and packet error rate in [6]). A drawback of these schemes is that they do not consider ATR, the channel availability.

Reference [7] proposed to balance effective channel busy-time (i.e., time the channel is busy for successful transmissions) among cells. Because it does not discriminate the time corresponding to unsuccessful transmissions and the idle time, [7] may suffer from such estimation errors. Moreover, because [7] ignores offered traffic volume, it might take a long time until load is balanced.

ATR is also the ratio of the amount of bandwidth consumed for transmissions to the total bandwidth. Thus, an increase of the offered traffic volume (injected traffic) increases ATR until it reaches its saturation value. Further increase of the offered traffic, however, results in congestion (i.e., buffer overflow) that significantly hampers communication performance. Since a cell with a small ATR can accommodate additional traffic, the overall throughput can be improved by moving STAs from the congested cell to such a non-congested cell. References [8], [4] proposed schemes that balance traffic volume among cells. A drawback of the schemes is that they do not consider the fact that the overlapping cells operating under the same frequency channel share the same channel resource. Moreover, [7], [8], [4] do not consider the link quality (the second component of (3)), and thus they may force STAs to use links with poor quality, degrading the users' throughput. To the best of our knowledge, none of the existing schemes takes account of the overall of (3), thus they do not necessarily maximize throughput.

To this end, we propose an association control scheme that takes account of the overall of (3). The direct target of the proposed scheme is to maximize the sum of the multiplication of the second and third components of (3), r and ATR, by taking account of RSSI, ATR, and the offered traffic volume. The success probability, P_{sc}, is indirectly maximized by maintaining ATR smaller than a threshold value. The next section provides an analytical study on why and how P_{sc} can be maximized by controlling ATR. An important difference of the proposed scheme from the previous schemes is that because both the offered traffic and channel availability are estimated for each cell, the proposed scheme does not aim at load balancing. This enables the scheme to maximize throughput by utilizing fewer APs, providing a positive impact on energy efficiency.

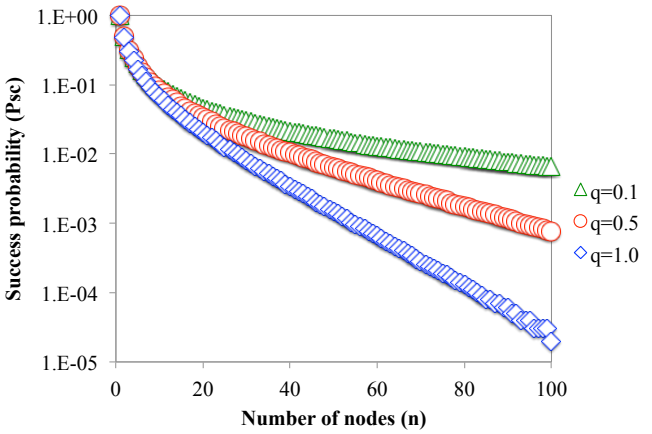


Figure 2. Success probability as a function of the number of contending nodes. The figure shows that P_{sc} sharply decreases with the increase of the number of nodes. The smaller the probability of having a pending frame (q), the larger is the P_{sc}.

III. Study on Air-Time Ratio and Success Probability

Equation (3) shows that the throughput of a node can be expressed by the multiplication of the three components: the success probability (P_{sc}), the average data transmission rate (r), and air-time ratio (ATR). The first and second components are further expressed as functions of the number of contending nodes (n) and link quality (RSSI), respectively. RSSI and ATR are local link and channel condition that can easily be monitored using the existing WLAN devices. On the other hand, the nodes that contribute P_{sc} can be beyond the transmission range of each other (i.e., they cannot correctly receive the packets from each other). Therefore it is an extremely difficult task to count the number of such nodes.

This section will show that the success probability (P_{sc}) can be indirectly maximized by controlling the air-time ratio (ATR). Before getting into the details, we will first show that it is necessary to maximize P_{sc} for throughput maximization.

As (2) shows, P_{sc} is a function of the number of nodes and the channel access probability (τ). τ is further expressed in [10]

$$\tau = q \times \frac{1 + \sum_{k=1}^{L_{retry}} (1 - P_{sc})^k}{CW_0[1 + \sum_{k=1}^{L_{retry}} 2^k (1 - P_{sc})^k]} \quad , \qquad (5)$$

where q is the probability of having a pending data frame (to transmit), CW_0 is the minimum contention window size, and L_{retry} is the frame retransmission limit [11]. We now calculate the success probability (using (5) and (2)) for the IEEE 802.11a system [11]. Table I shows the parameters used in the calculation. Fig. 2 shows the success probability (P_{sc})

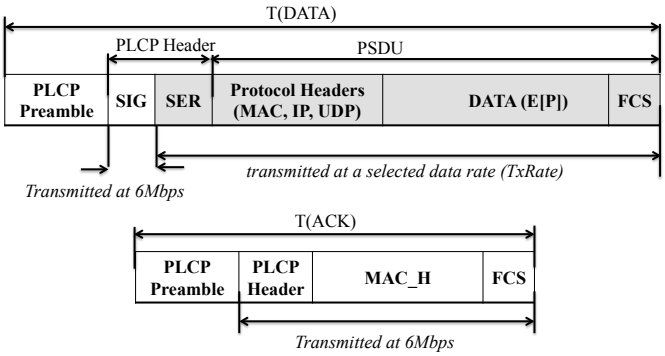


Figure 3. T' the sum of the times required for transmissions of a data frame and the corresponding acknowledgment frame. PLCP SIGNAL field duration and ACK frame are generally coded with 1/2 BPSK (i.e., 6 Mbps rate) in the IEEE 802.11a systems. The shaded area of the data frame is transmitted at a rate (TxRate) selected by the MAC rate adaptation mechanism.

TABLE I
PARAMETERS USED IN THE PERFORMANCE ANALYSIS

Parameter	Value	Note
σ	$9\mu s$	Slot time
SIFS	$16\mu s$	Short Inter-Frame Space
DIFS	$34\mu s$	DCF Inter-Frame Space
CW_0	16	Minimum contention window size
L_{retry}	4	Frame retransmission limit
PLCP preamble	$16\mu s$	PLCP preamble duration
PLCP_SIG	3 bytes	PLCP SIGNAL field length
PLCP_SER	4 bytes	PLCP SERVICE field length
MAC_H of data frame	24 bytes	MAC header of data frame
MAC_H of ACK frame	10 bytes	MAC header of ACK frame
FCS field	2 bytes	
IP header size	40 bytes	IPv6 header
UDP header size	8 bytes	
E[P]	500 bytes	Data size

for different numbers of nodes. While the smaller the q, the larger is the P_{sc}, it is clear that P_{sc} sharply degrades with the increase of the number of nodes (note that the vertical axis is on a logarithmic scale).

We now express ATR by P_{sc}. P_{tr}, the probability that there is at least one station transmits (see (1)) is

$$P_{tr} = 1 - (1-\tau)^n \quad . \tag{6}$$

By substituting P_{tr} to the formula of ATR (the 3rd component of (3)), the following relation of ATR and P_{sc} can be achieved.

$$ATR = \frac{1}{P_{sc}\frac{(1-\tau)\sigma}{T'\tau} + \frac{DIFS+SIFS}{T'} + 1} \tag{7}$$

Here T' is the time required for the transmission of a data frame without including the inter-frame spaces (DIFS and SIFS). Because T' is more convenient to monitor than T (which includes the inter-frame spaces) in the real systems, we use T' in this and the following sessions. For simplification of the mathematical calculations, we target the basic access method (i.e., the RTS/CTS (Request to Send and Clear to Send) handshake is not used). In such a case, T' is the sum of the average time required for the transmissions of a data frame and the corresponding acknowledgement frame (ACK) (see Fig. 3).

Substituting (7) to (3), the throughput of a node is expressed slightly differently from (3):

$$s = r' \times \frac{1}{\frac{(1-\tau)\sigma}{T'\tau} + (\frac{DIFS+SIFS}{T'} + 1)\frac{1}{P_{sc}}} \quad , \tag{8}$$

where r'=E[P]/T'.

Using (8), (5), and (2) we now calculate the throughput for the IEEE 802.11a system. In the IEEE 802.11a, the acknowledgement frame (ACK) and the SIGNAL field of the PLCP header (the PLCP header fields excluding the SERVICE field) are generally coded with 1/2 BPSK (6 Mbps). The PSDU (physical layer service data unit) and the SERVICE field of the PLCP header in the data frame (the shaded part in Fig. 3)

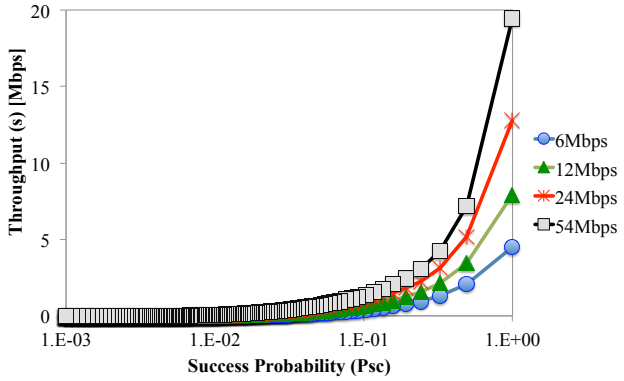


Figure 4. The throughput performance as a function of the success probability. The figure shows that for a sufficient throughput is achieved only when the success probability is large.

are subject to the MAC rate adaptation and thus we calculate the corresponding transmission time for different data rates. Fig. 4 shows the impact of the success probability, P_{sc}, on the throughput performance for different data rates. First of all, the figure shows that higher the utilized data rate (i.e., better link quality), the larger is the achieved throughout, implying the necessity of considering the link quality, i.e., the second component of (3). An important observation that can be made from the figure is that a sufficient throughput can be achieved only when P_{sc} is large.

Now we will show the impact of P_{sc} on ATR. In fact we can easily see in (7) that ATR takes on its maximum value, ATR_{max}, when P_{sc} is 0. In this case, however, the channel is occupied with packet collisions, and obviously, the throughput is minimized (see Figure 4). On the other hand, if P_{sc} is 1 (i.e., channel utilization is maximized), ATR takes on a value smaller than ATR_{max}, implying that the optimal value of ATR, ATR_{opt} is below ATR_{max} [12].

Now we will have a closer look at the relationship of ATR and P_{sc}. We substitute (5) to (7) and calculate ATR for the IEEE 802.11a system.

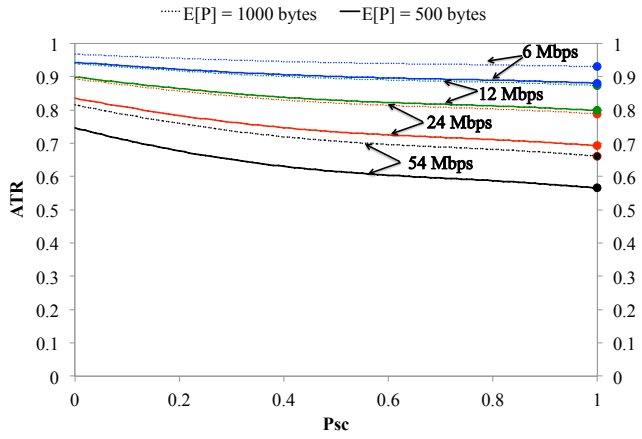


Figure 5. ATR characteristics with respect to the success probability, P_{sc}. Note that the solid (dashed) lines correspond to 500 bytes (1000 bytes) of data size. The figure clearly shows that ATR is maximized when P_{sc} is minimized. The optimal value of ATR (that maximizes P_{sc}) is smaller than the maximum value of ATR. The figure also shows that the larger the data rate, the smaller is the ATR and the larger the data size, the larger is the ATR.

Fig. 5 shows the ATR characteristics with respect to P_{sc} for different data sizes (E[P]) and data rates. The figure clearly shows that ATR is maximized when P_{sc} is minimized; an increase of P_{sc} results in a gradual decrease of ATR. Moreover the larger the utilized data rate, the smaller is the ATR; the larger the data size, the larger is the ATR. The figure clearly shows that different optimal values exist for different data rate and data size. For example ATR_{opt} is 0.57, 0.66, 0.88, and 0.92 for the data rate and data size combinations of (54 Mbps, 500 bytes), (54 Mbps, 1000 bytes), (6 Mbps, 500 bytes), and (6 Mbps, 1000 bytes), respectively. (ATR_{opt} are shown with the round markers in the figure).

Our objective is to find a threshold value, ATR_{th}, which is used by our association control scheme to ensure P_{sc} be sufficiently large. Because ATR_{opt} takes on different values depending on the data size and the data rate, it is maybe possible to design a mechanism that finds the exact ATR_{opt} and set ATR_{th} to that value. However, designing such a mechanism can be very complex because it requires a precise estimation of the average data rate and the average data size. We believe that it is simple and thus realistic to set ATR_{th} to a fixed value. In WLANs, the users can be very close to the APs, i.e., the link quality (RSSI) can be good enough for a use of the highest data rate. Thus, it is safe to set ATR_{th} to ATR_{opt} corresponding to the highest data rate (i.e., 54 Mbps for the IEEE 802.11a). The data size, on the other hand, can be monitored or it can be set to a value, which is used by the typical applications. Thus according to Fig. 5, if the data size is 500 bytes, ATR_{th} should be set to 0.57. In fact, in our previous work [1], we have empirically found that it is appropriate to set ATR_{th} to 0.58 for WLANs for scenarios where data size takes on a value between 500 to 540 bytes. The analytical calculation achieved in Fig. 5 proves the correctness of the empirically-found threshold value.

As the success probability, P_{sc}, can be maximized by maintaining ATR below the threshold value (ATR_{th}), the throughput maximization problem can be rewritten as

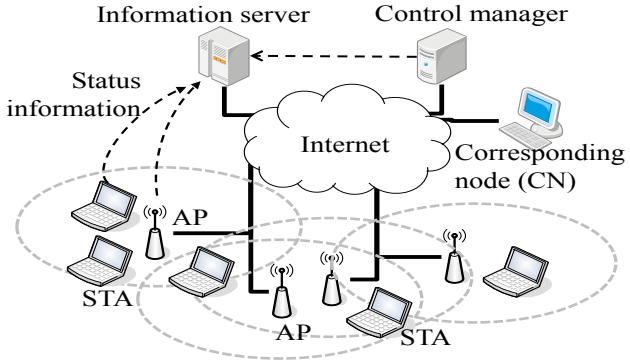


Figure 6. Besides having user-owned APs and STAs, the target community network has a centralized control manager. The information that are necessary for network management is collected by APs and/or STAs and gathered at an information server.

$$\begin{array}{ll} \text{Maximize} & S = \sum_{i=1}^{N} \sum_{j=1}^{n_i} [r_{i,j} \times ATR_{i,j}] \\ \text{subject to} & ATR_{i,j} < ATR_{th} \quad \forall (i, j) \end{array} . \quad (9)$$

where N is the number of cells, n_i is the number of STAs at the i^{th} cell. $r_{i,j}$ and $ATR_{i,j}$ are the average frame transmission rate and ATR corresponds to the j^{th} node at the i^{th} cell.

IV. Proposed Scheme

A user-owned AP is integrated into a community network by a common equipment/software program provided by the entity (e.g., FON [2]). Besides integrating users' APs, the entity can play an important role for, e.g., network management for better communication quality. Hence, community networks are centrally controllable and we expect that the community members (i.e., the users) are cooperative to such a control. To this end, while a distributed mechanism can also be designed, in this paper, we propose a centralized association control scheme due to ease of management and better performance [7], [8], [5]. The proposed scheme aims at maximizing network throughput by a congestion alleviation mechanism and reducing energy consumption by a cell aggregation mechanism. For a large community network, the network can be divided into sub-networks, which are independently and separately controlled. Fig. 6 shows the network architecture.

Besides APs and STAs, the network has an information server (server) and control manager (manager). The server and manager can be physically separated or coexist in a same machine. APs and STAs periodically inform the server of information on the link quality, channel availability, and traffic volume. Periodically referring to the information, the manager triggers STAs' handover for improved network throughput and/or energy efficiency.

A. Estimation of Channel Availability and Offered Traffic

STAs and APs measure RSSI, ATR, and the offered traffic volume, and inform the server of the information. The manager

uses the information to estimate channel availability and traffic condition for each cell, and changes STAs associations.

- Frame transmission rate

By periodically performing channel scanning, STAs are aware of the existence of the neighboring APs and the corresponding $RSSI_{AP,STA}$. Such background scanning is supported by the off-the-shelf wireless LAN cards [13], and some efforts have been made for fast channel scanning [14]. Transmission rate for frame payload field is estimated from $RSSI_{AP,STA}$ and finally the frame transmission rate, $r_{AP,STA}$ ($\approx$E[P]/T), for each pair of STA and AP is calculated.

- ATR

Because ATR value does not largely change over the space (e.g., it is around the same in the transmission range), only APs measure ATR on their operating channel. Such a measurement can easily be made using the existing WLAN cards [7], [5], [15]. We empirically found that the appropriate values for ATR_{th} are 0.58.

- Potential Throughput

The manager estimates the maximum achievable throughput for each pair of STA and its neighboring AP (which is not the STAs currently associated AP). Let $PT_{STA,AP}$ (potential throughput) represent the maximum achievable throughput for such a STA and an AP. Equation (3) shows that the potential throughput is a function of P_{sc}, the estimated transmission rate ($r_{AP,STA}$), and ATR_{AP}. As discussed in the previous section, however, P_{sc} can be maximized by ensuring ATR below ATR_{th}. This enables PT to be estimated only from $r_{AP,STA}$ and ATR_{AP}. Thus the manager calculates PT for each STA and its neighboring AP as

$$PT_{AP,STA}=\begin{cases} 0 & ATR_{AP}\geq ATR_{th} \\ (ATR_{th}-ATR_{AP})\times r_{AP,STA} & \text{otherwise} \end{cases}. \quad (10)$$

The upper equation is to not move the STA to the AP, at which ATR_{AP} exceeds ATR_{th}. Otherwise, the AP is a candidate destination AP for the STA, and the maximum achievable throughput at the candidate cell is expressed as the lower equation. In our previous work [15], we confirmed that such an estimation of PT can be achieved with a high accuracy using the existing WLAN cards.

- Offered Traffic

A STA can be moved to a neighboring AP, if it does not cause congestion at the neighboring cell. The condition can be verified by comparing the offered traffic volume for the STA and $PT_{STA,AP}$ (to be discussed later). Letting $EnqueueRate_{A,B}$ be the rate at which packets destined to node B are inserted into the IP queue at node A, the offered traffic volume for a STA is defined as

$$OfferedRate_{STA}=EnqueueRate_{STA,AP}+EnqueueRate_{AP,STA}. \quad (11)$$

If the WLAN is the bottleneck link of the end-to-end route, the OfferedRate is approximately equal to the traffic generation rate.

B. Congestion Alleviation

A cell is considered to be congested if

$$ATR>ATR_{th} \quad . \quad (12)$$

If the condition (12) is met for a cell, the manager checks if the aggregate offered rate exceeds the aggregate packet throughput for that cell, i.e.,

$$\alpha \sum_{STAs} OfferedRate_{STA} > \sum_{STAs} PacketThroughput_{STA} \quad (13)$$

Here α ($<$1) is a system parameter to absorb rate fluctuation. $PacketThroughput_{STA}$ is the sum of the rates at which packets are successfully transmitted on the uplink and downlink for the STA. The condition (13) indicates that one or more nodes belonging to the cell are experiencing buffer overflow. It is possible that a cell satisfies (12) but not (13), in a case, when the cell does not have much traffic but the channel is congested due to the transmission activities at the overlapping cells.

A cell that satisfies both (12) and (13) becomes a target cell of congestion control. Letting APt be the AP of the target cell, the manager selects STAs to move from APt to its neighboring cells. The association control is made based on the following policies:

1) The number of moving STAs should be as small as possible.
2) A STA should be moved only if it will not cause congestion at the destination cell.
3) Among the candidate destination APs, the STA should be moved to the AP corresponding to the strongest RSSI.

The more the number of handovers, the larger is the control overhead. Thus policy 1 is to minimize the number of moving STAs. To support this objective, we define load of a STA, as $Load_{STA}$=$OfferedRate_{STA}$/$TxRate_{STA,APt}$. Here $TxRate_{STA,APt}$ is the data rate used for transmissions of frame payload fields between the STA and APt (the shaded part in Fig. 3). Obviously, the larger the offered traffic and/or the lower the transmission rate, the heavier loaded is the STA for APt. For policy 1, heavier loaded STAs are preferred to be moved (from APt) over lighter loaded STAs. For policy 2, a STA is moved to a neighboring AP, APd, only if the condition (14) is met.

$$OfferedRate_{STA} < PT_{STA,APd} \quad (14)$$

In other words, the STA is moved to APd only if the destination cell can accommodate the offered traffic volume for the STA. Finally, among the candidate destination APs (i.e., the APs that satisfy (14) for the STA), the AP corresponding to the strongest RSSI is selected as the destination AP for the STA.

When a STA, STAm, is selected to be moved from APt, the manager updates the aggregate offered rate (see (13)) for the target cell by decrementing it by $OfferedRate_{STAm}$. Moreover ATR for the destination cell and its overlapping cells (which operate using the same channel) is incremented by $OfferedRate_{STAm}/r_{APd,STAm}$. After updating the values, the

manager checks if the target cell still satisfies (13). If it does, the manager searches the next STA to move from APt.

- Discussion on TCP traffic

TCP reacts to congestion and controls its rate based on AIMD (Additive Increase Multiplicative Decrease) algorithm. However such a rate change occurs in the order of RTT (round-trip time), i.e., in order of milliseconds, while the control period at the manager is in order of seconds. Therefore we expect that the proposed scheme does not react to the AIMD-based rate fluctuation, but only to the gradual change of the average rate. Hence, the proposed scheme and TCP can stably coexist. Moreover because TCP adjusts its traffic rate, $\text{OfferedRate}_{\text{STA}}$ for STA might be largely changed due to the STAs movement. However, it should be reminded that a STA is moved to a neighboring cell only if the destination cell can accommodate the current OfferedRate for the STA (see (14)). Thus, we expect that TCP throughput will be increased or maintained after the STAs movement. Furthermore, since some amount of channel resource is released at the previous cell of the moving STA, STAs in that cell can now increase their rate.

C. Cell Aggregation

Since both of the offered traffic volume and the channel availability are known for each cell, load balancing among cells is not necessary. Indeed, if all the associated STAs of an AP can be moved to its neighboring cells without hampering the network throughput, such STA movements should be encouraged for energy efficiency, since the AP can now be in the power-saving mode. Our scheme can provide such an association control based on the following policies:

1) The target AP is an AP that is associated with preferably few STAs, which all can be moved to the neighboring cells.
2) To suppress channel interference, the target AP should have overlapping cells that operate using the same frequency channel.
3) Policy 2 described in the previous subsection.
4) Policy 3 described in the previous subsection.

The manager triggers a handover only if destination APs are found for all the STAs of the target AP. A detailed protocol design for actually putting APs in the power-saving mode is left for our future work. The main concern of such a protocol design is to ensure newly arriving STAs are covered by the network. For such a control, Wake-on-WLAN technology [16] can be used.

D. Changing STA's Association

To change a STA's association, the manager sends a control frame to the STA, indicating the destination AP and the corresponding channel information. Such a network directed association control can be supported by the upcoming IEEE 802.11v [17], which enables APs to explicitly request STAs to re-associate with an alternate AP.

TABLE II
USER DISTRIBUTION FOR EACH SCENARIO

	Circle1 (station)	Circle2 (condominium)	Circle3 (McDonalds)	Circle4 (overall area)
S1	0	0	0	40
S2	10	10	10	10
S3	10	20	0	10
S4	10	20	10	0
S5	20	10	0	10
S6	30	0	0	10
S7	40	0	0	0

V. SIMULATION EVALUATION

Using Scenargie network simulator [18], we investigate the proposed scheme with and without cell aggregation capability. The performance of the proposed scheme is compared against:

- Strongest RSSI: The traditional AP selection scheme.
- LB(NumofSTAs): A load balancing scheme [3], which takes account of the number of STAs and RSSI.
- LB(Traffic): A load balancing scheme [8], where load of a cell is defined by traffic rate.

In each scheme, STAs initially associate with APs based on the strongest RSSI policy. In the proposed scheme, STAs inform the server of the measured information using a 160-bytes packet. The manager checks the collected information in every second. 20-bytes of packets are used for handover requests and replies between the manager and moving STAs. α (see (13)) is set to 0.98.

The performances of the schemes are investigated targeting the real-world scenario depicted in Fig. 1, where 14 APs are non-uniformly distributed in a 600×600 m^2 area. 8 APs are concentrated in the small area (around the condominium), the remaining 6 APs are installed in the rest of the overall area as shown in the figure. The network operates using the IEEE 802.11a [11], using 3 orthogonal frequency channels. The frequency channels are allocated (to the APs) following a simple graph coloring technique.

STAs (users) can be anywhere, but it is natural to expect that users are especially attracted to the public places specifically, the train station, condominium building (for the café), and McDonalds shown in the target area. Thus, in our simulations, 40 STAs are distributed in the circle-shaped areas centered at the 1) train station, 2) condominium building, 3) McDonalds, and 4) the center of the map, with a radius of 100, 10, 10 and 300 meters, respectively (see Fig. 1). The first three areas are to create users concentration in the public places, while the last area is for uniform distribution of users in the overall area. Table II shows the simulation scenarios, which have different number of STAs in each circle-shaped area. The smaller the scenario number, the more uniform is the STAs' distribution.

The simulation evaluations are conducted for TCP and UDP traffic. In TCP simulations, STAs upload 5 MB of file using FTP/TCP-SACK. In UDP simulations, STAs have uplink and downlink CBR traffic generated at a random rate in the range of [0 Mbps, 1.2 Mbps]. The data size is set to 512 bytes for

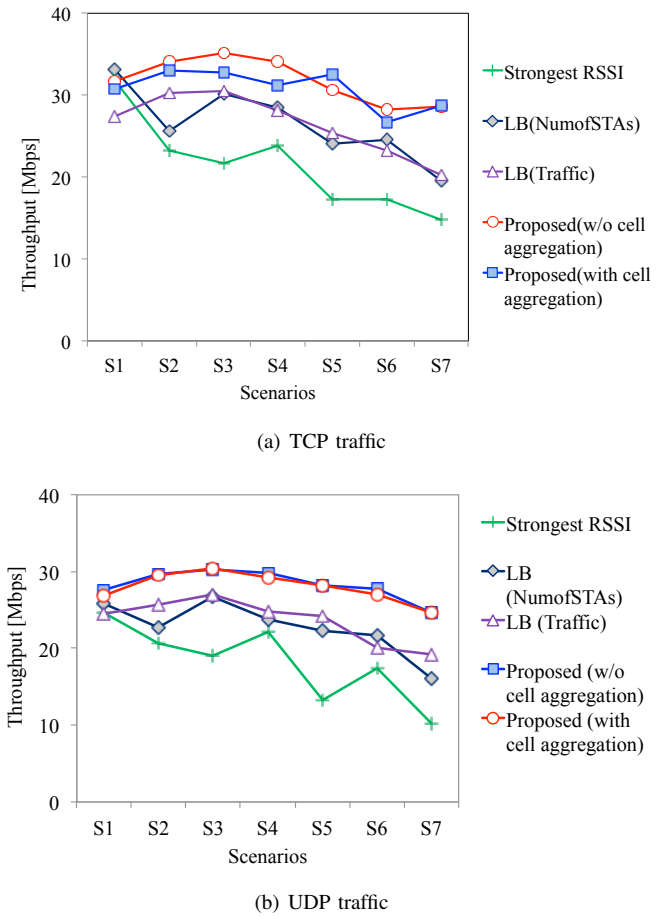


Figure 7. Comparing throughput performance for TCP and UDP traffic. The figures show that Strongest RSSI scheme is inferior to the remaining schemes and the proposed scheme shows the best performance.

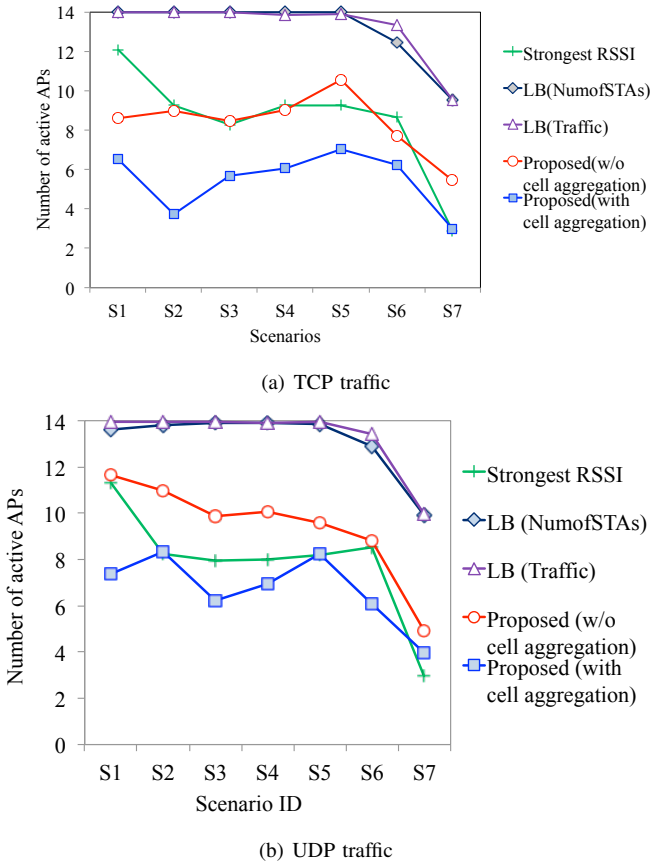


Figure 8. Comparing the number of active APs for each scheme. The figures show that the load balancing schemes tend to use all the existing APs. The proposed scheme without cell aggregation utilizes around the same number of APs as that of Strongest RSSI scheme. The proposed scheme with cell aggregation utilizes the smallest number of APs.

both TCP and UDP traffic.

Fig. 7 compares the network throughput for the TCP and UDP traffic. As the figures show, the performance of Strongest RSSI scheme is inferior to the remaining schemes and the proposed scheme shows the best performance. The proposed scheme with the cell aggregation mechanism achieves around the same throughput performance as the scheme without cell aggregation, especially for UDP traffic. The load balancing schemes have lower throughput than the proposed scheme, conceivably because they do not take account of the overall of (3).

The figures also show that, in general, the throughput tends to be smaller when STAs' distribution is less uniform (e.g., the throughput of S7 is smaller than that of S5). As discussed in Section II, the reason is clear for the strongest RSSI scheme (many STAs select the same AP). The reason for the remaining schemes is as follows. When STAs are highly concentrated around a particular AP, the schemes have to move some of the STAs to farther APs. This reduces the frame transmission rate, r, for the moving STAs, resulting in lower throughput compared to that of a scenario where STAs' distribution is more uniform. Nevertheless, compared with the strongest RSSI scheme, the overall throughput is improved due to an increase of the 1st and 3rd components of (3). Finally, S1 (where STAs' distribution is uniform), however, does not show the largest throughput due to the non-uniform AP distribution.

Fig. 8 compares the number of active APs, i.e., the number of APs that actually serve for the users. As the figure shows, the load balancing schemes use all the existing APs (except in scenario S7, where STAs are concentrated only at the train station). Strongest RSSI scheme, on the other hand, does not use many APs due to its simple AP selection policy. The proposed scheme without cell aggregation utilizes around the same number of APs as that of Strongest RSSI scheme. Finally, the scheme with cell aggregation utilizes the smallest number of APs. This is especially attractive because compared to the existing schemes, the proposed scheme largely improves throughput by utilizing fewer APs. Our future work includes a study on how much energy reduction can be achieved by such an association control.

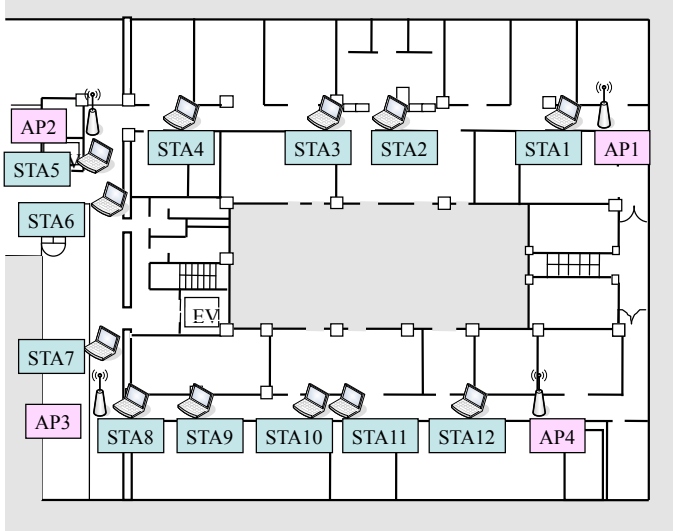


Figure 9. An active office area, which has a complicated room and obstacle structure and intermittent movement of people, is used as the test environment. The test scenario consists of 4 APs and 12 STAs. STAs are initially associated with the APs using the strongest RSSI method.

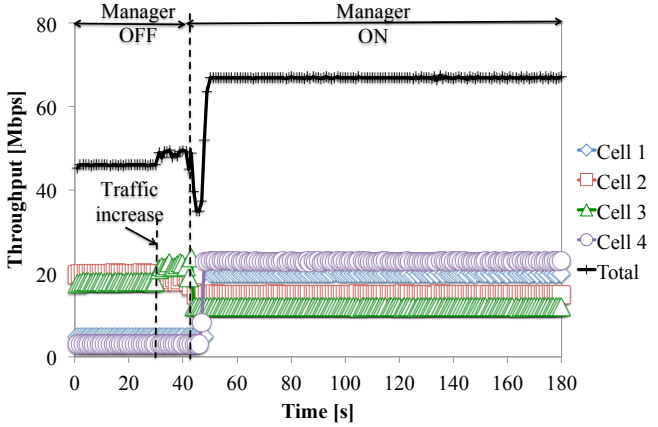


Figure 10. Throughput performance of the congestion alleviation method. In this scenario all STAs have CBR/UDP traffic. Upon activation of the manager, STAs are moved from the congested cells (cells 2 and 3) to the non-congested neighboring cells (cells 1 and 4). The aggregate throughput is largely improved by the association control.

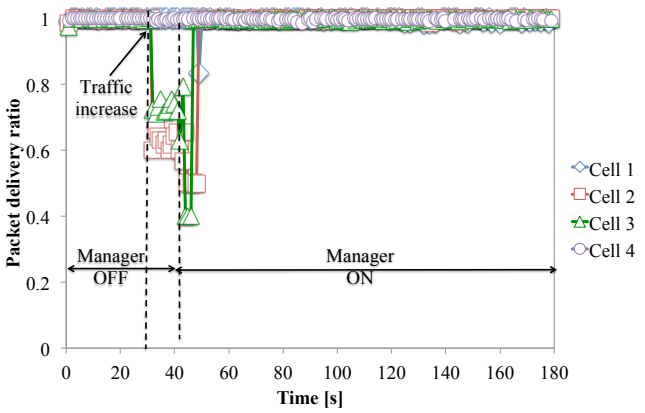


Figure 11. Packet delivery performance of the congestion alleviation method. Due to the additional traffic injected after around 30 seconds, the packet delivery ratio largely degrades down to 60%. Upon activation of the manager, the delivery performance is recovered back to 100%.

VI. Testbed Evaluation

We implemented the proposed scheme in a wireless testbed and evaluated its throughput performance. Computers with Cent OS 5.5 (kernel 2.6.25-17) are used for STAs, APs, and the manager (note that the manager and server coexist on a same machine). The APs, the manager, and a source computer (acting as a corresponding node (CN)) are connected to a 100 Mbps Ethernet. The APs and STAs are equipped with IEEE 802.11a WLAN cards made by NEC (Aterm WL54AG). We modified the Atheros device driver to enable measurements of ATR and PacketThroughput (see (13)). The packet transmission rate from the kernel to the device driver is monitored to measure EnqueueRate (see (11)). TCP is used for information collection at the server and for handover requests and replies.

As the testbed environment, we selected an active office area, which has a complicated room and obstacle structure and intermittent movement of people. Fig 9 shows the testbed environment. It consists of multiple rooms connected by long corridors, surrounding a central well. Unless otherwise noted, the test system consists 4 access points (AP1-AP4) and 12 STAs (STA1-STA12) that are positioned as illustrated in the figure. STAs are initially associated with APs based on the strongest RSSI method. The relative positioning enables each AP to be initially associated with 3 STAs (STA1-STA3 are with AP1, STA4-STA6 are with AP2, STA7-STA9 are with AP3, and STA10-STA12 are with AP4).

To see the impact of the proposed scheme, the manager is initially disabled and it is activated after around 40 seconds. Unless otherwise noted, the offered traffic volume for individual cells are initially set to be largely different. Moreover, in order to create traffic dynamics, additional traffic is injected to some cells after around 30 seconds. To simplify experimentation, the CN (corresponding node) generates traffic for individual STAs (i.e., downlink traffic), enabling the offered traffic volume to be adjusted only at CN. Three orthogonal frequency channels of the IEEE 802.11a are allocated such that AP1 and AP3 share the same channel and AP2 and AP4 use the remaining two channels. The testbed evaluations made for three scenarios: the first scenario is to evaluate congestion alleviation method for UDP traffic, the second scenario is to evaluate congestion alleviation method for TCP and UDP traffic, and the last scenario is to evaluate cell aggregation method.

A. Congestion alleviation for UDP traffic

CBR/UDP traffic is generated for individual STAs. Initial setting of the aggregate offered traffic at cells 1, 2, 3 and 4 are 5, 20, 18, and 3 Mbps, respectively. At 30 seconds, the offered traffic rate for cells 2 and 3 are further increased by 10 Mbps.

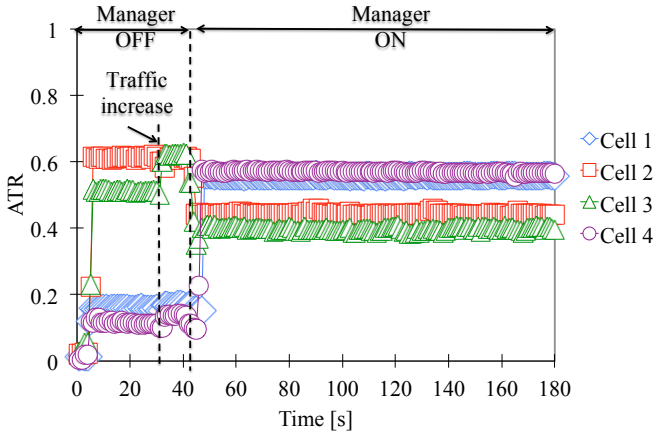


Figure 12. ATR measured at each APs. The figure shows that ATR can be well tuned by the association control to the desired value.

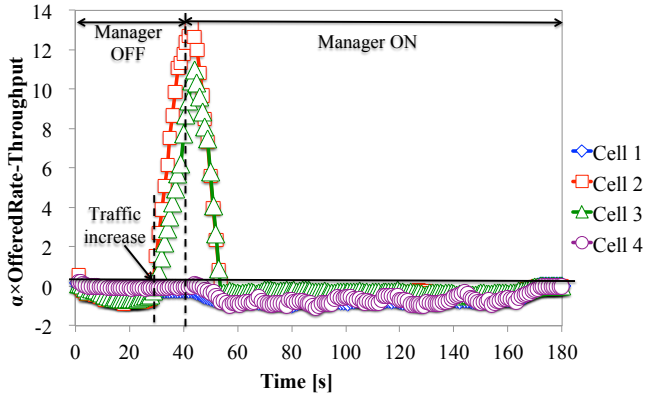


Figure 13. $\alpha \times \sum_{STAs} OfferedRate_{STAs} - \sum_{STAs} PacketThroughput_{STAs}$ at each cell. The figure shows that the parameter reacts well to buffer overflow.

Figs. 10 and 11 show the time series plots of the throughput and packet delivery performance for each cell. Because cells 2 and 3 were initially very loaded, the additional traffic injected to the cells (at around 30 seconds) results in only a slight increase of the total throughput. As the offered traffic volume requires a larger resource, as Fig. 11 shows, the cells could not accommodate the overall offered traffic, and hence the packet delivery ratio drops down to 60%. Upon activation of the manager (at around 40 seconds), STAs are moved from congested cells (cells 2 and 3) to non-congested neighboring cells (cells 1 and 4), and the packet delivery ratio is maximized by the association control.

Figs. 12 and 13 show the time series plots of ATR and $\alpha \times \sum_{STAs} OfferedRate_{STAs} - \sum_{STAs} PacketThroughput_{STAs}$, the parameters used to trigger congestion control (see (12) and (13)). Two important observations can be made from Fig. 12. Firstly, before the congestion (before 30 seconds), ATR at Cell2 already exceeds the threshold value, but as it can be seen in Figs. 11 and 13, there was no buffer overflow. This teaches us that ATR above ATR_{th} does not necessarily indicates buffer overflow. The second point can be observed

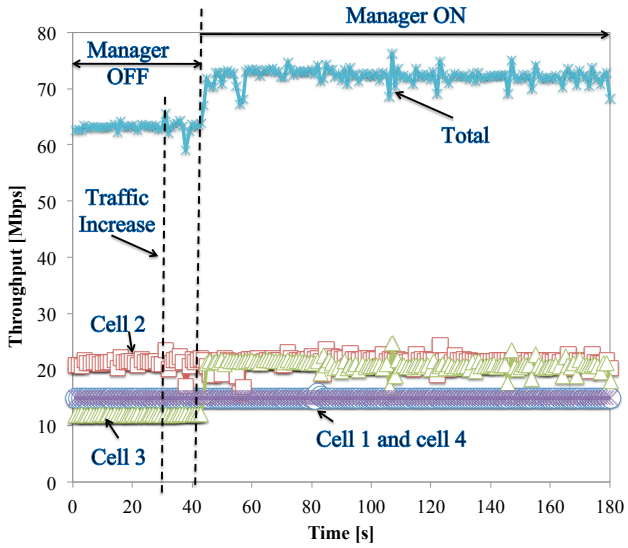


Figure 14. Throughput performance of each cell. In this scenario, TCP and UDP traffic share the same resource at Cell2. Upon activation of the manager, a STA with UDP traffic is moved to Cell3. With this association control, more than 10 Mbps throughput improvement is achieved.

in Fig. 12 is that ATR can be tuned well to the desired level, ATR_{th}. Fig. 13 shows that $\alpha \times \sum_{STAs} OfferedRate_{STAs} - \sum_{STAs} PacketThroughput_{STAs}$ reacts well to the congestion and together with ATR, it can be a good indication of congestion for UDP traffic.

Figs. 10-13 show that unfortunately, it took approximately 6 seconds to complete the handover (without depending on the number of moving STAs). The 6 seconds are used for 1) MAC layer disassociation/association, 2) IP route advertisement, 3) IP duplicate address detection, 4) Mobile IP binding update, and 5) Mobile IP binding acknowledgement. Among the operations, there was a software bug corresponding to 2) and we confirmed that by fixing this bug, handover time can be reduced down to 3 seconds. We are now working on this issue.

B. Congestion control for TCP and UDP traffic

The CN generates TCP traffic to the STAs at cells 1, 3, and 4. For Cell2, TCP traffic is generated for STA5, and UDP traffic is generated for STA4 and STA6. The initial settings of the maximum offered traffic volume (i.e., the sum of the CBR rate and the maximum rate of TCP) for the cells 1, 2, 3, and 4 are 15 , 27, 12, and 15 Mbps, respectively. At around 30 seconds, the UDP traffic for STA4 is further increased by 5 Mbps. Fig. 14 shows the time series plots of the aggregate throughput and throughput of individual cells. Upon activation of the manager, STA6 is moved from Cell2 to Cell3. This improves the aggregate throughput by approximately 10 Mbps.

Fig. 15 gives a closer look at the time series plots of the individual throughput at Cell2 and Cell3. The offered CBR rate and the maximum offered TCP rate for each STA are also noted in the figure. Until 30 seconds, the UDP throughput achieved for STA4 and STA6 (in Cell2) are 10 Mbps, which are equal to the offered rate. On the other hand, the throughput

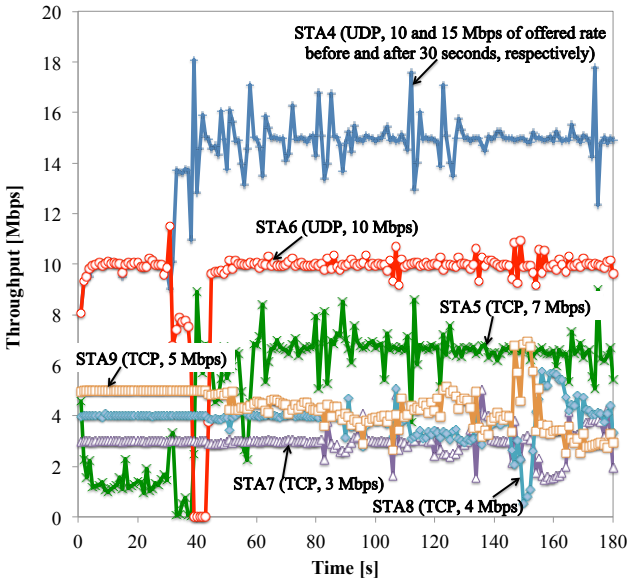


Figure 15. Throughput of individual STAs at cells 2 and 3. Upon activation of the manager, STA6 is moved from Cell2 to Cell3. This enables TCP traffic at STA5 to increase its rate up to its maximum rate and the UDP traffic at STA4 also achieves its maximum throughput.

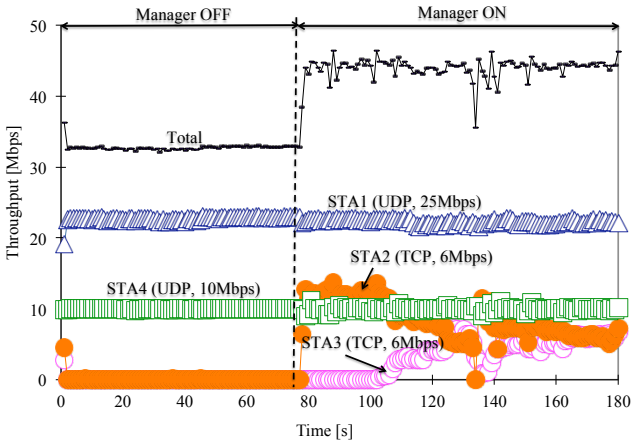


Figure 16. Thoughput of individual STAs for a scenario where STAs with TCP traffic are moved (STA2 and STA3). The TCP traffic could achieve increased throughput at the destination cell.

of TCP traffic (for STA5) is only 2 Mbps, which is much lower than the maximum offered rate (7 Mbps). The additional traffic injected for STA4 (at around 30 seconds) results in a large throughput degradation for both STA6 and STA5. The throughput degradation is especially significant for the TCP traffic for STA5 (down to 0 Mbps). This clearly shows severe unfairness between TCP and UDP traffic. At around 40 seconds, the manager is activated, then STA6 is moved to cell 3. After approximately 5 seconds of handover delay (due to the issues mentioned in the previous subsection), all the traffic at both the cells could be transmitted at their maximum rate, resulting in 10 Mbps of throughput improvement.

In the above mentioned scenario, a STA with UDP traffic is moved from the congested cell. Fig. 16 shows results achieved from an experiment where STAs with TCP traffic are moved. In the experiment, the target scenario consists of two cells: one has three STAs (STA1-STA3) and the other has one STA (STA4). The manager is activated at around 70 seconds. STA1 and STA4 have UDP traffic with 25 and 10 Mbps of offered rates, respectively. STA2 and STA3 have TCP traffic with maximum 6 Mbps of offered rate. As the figure shows, until activation of the manager, due to the large traffic volume offered to STA1, the coexisting TCP traffic significantly reduced their traffic rate down to 0 Mbps. Upon activation of the manager, the STAs with TCP traffic are moved from Cell1 to Cell2. The association control results in the TCPs increase their traffic rate at the destination cell.

The results achieved in Figs. 10-16 also show that monitoring/estimation of traffic, channel, and link condition made at Layer1-Layer 3 are realized with sufficient accuracy. However because TCP controls its rate at Layer 4, the OfferedTraffic estimated at Layer 3 is the result of the TCP's rate adaptation. This implies, on the one hand, that the proposed congestion alleviation scheme does not negatively affect the TCP's congestion control, and the two congestion control schemes can stably coexist. On the other hand, however, if the proposed scheme is aware of the TCP's maximum rate, further throughput improvement is possible. The following is a list of some approaches for this challenging task.

- Enhancement of the proposed scheme by tuning the control parameter (e.g, by increasing α).
- Exploit TCP flow characteristics for association control. As Figs. 15 and 16 show, if TCP is transmitting at its maximum rate (i.e., at RWND), the throughput does not fluctuate with time. On the other hand, if TCP is not transmitting at its maximum rate, the throughput tends to fluctuate largely. By monitoring such characteristics, we believe that it is possible to "guess" that TCP is not transmitting at its maximum rate.
- Cross layer approach that enables direct information exchange between between Layer 4 and Layer 3.

C. Cell aggregation

We evaluate the performance of the cell aggregation technique. The testbed scenario is the same as shown in Fig. 9. Each cell has STAs with UDP traffic whose aggregate offered rate is 9 Mbps, thus congestion is not an issue. Fig. 17 shows the evaluation result. As the figure shows, upon activation of the manager (at around 30 seconds), the STAs at cells 1 and 4 are moved to cells 2 and 3, enabling AP1 and AP4 be in the power-saving mode. Although the throughput is degraded during the disconnection period, as soon as the STAs are associated with the destination APs, the network achieves its maximum throughput utilizing the half of the total number of APs.

VII. Conclusion

In community WLANs, APs and STAs (i.e., users) tend to concentrate on different areas. A concentration of STAs often

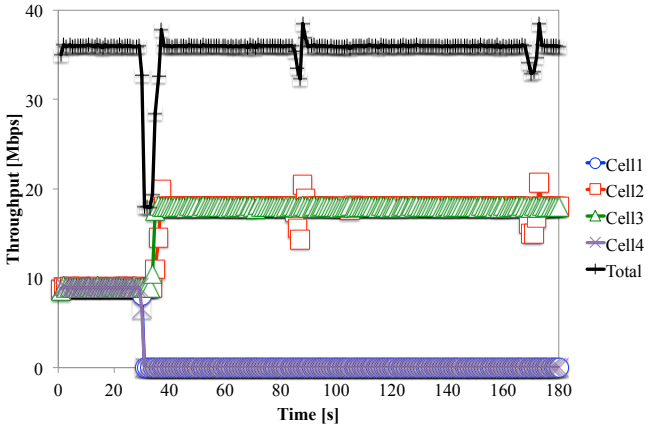


Figure 17. Throughput performance of the cell aggregation mechanism. STAs at cells 1 and 4 are moved to cells 2 and 3.

results in the APs and STAs in that particular area suffer from congestion. A concentration of APs, on the other hand, causes energy wastage. In this paper, we proposed an association control scheme that aims at maximizing throughput by congestion alleviation method and reducing energy consumption by cell aggregation method. We first analytically formulated that the throughput of a node belonging to a WLAN can be expressed as the multiplication of three components: success probability, frame transmission rate, and air-time ratio. The success probability is the probability of collision-free transmission. The frame transmission rate is the speed of the data transmission, and it is a function of the link quality. The air-time ratio is the ratio of the channel busy time to the total time. The frame transmission rate and air-time ratio can easily be monitored and estimated. The success probability, on the other hand, is a function of the number of contending nodes, which are extremely difficult to monitor in the real systems. Due to this reason, we extended our theoretical study and showed that success probability can be maximized by controlling air-time ratio. Finally, we proposed our association control scheme for throughput maximization and energy efficiency by taking account of the multiplication of frame transmission rate and air-time ratio. The performance of the proposed scheme is investigated by extensive evaluations using computer simulations and testbed experiments. Both the simulation and testbed evaluations strongly indicate the correctness of the theoretical work and the effectiveness of the proposed scheme for realistic network scenario with both UDP and TCP traffic. The testbed experiments prove that the proposed scheme and TCP can stably coexist for congestion control. We plan to further enhance the proposed scheme for improved TCP throughput. In addition to simulation and testbed evaluations, we intend to evaluate the scheme using statistical analysis. Our future work also includes a study on energy efficiency induced by the cell aggregation mechanism.

ACKNOWLEDGMENT

This research was performed under research contract of "Research and Development for Reliability Improvement by The Dynamic Utilization of Heterogeneous Radio Systems", for the Ministry of Internal Affairs and Communications, Japan.

REFERENCES

[1] O. Shagdar, S. Tang, A. Hasegawa, T. Shibata, and S. Obana. Association Control for Throughput Maximization and Energy Efficiency for Wireless LANs. In *Emerging 2011: Proceedings of the Third International Conference on Emerging Network Intelligence*, Lisbon, Portugal, November 2011. pp. 112-117.

[2] FON Community. http://corp.fon.com/ (last accessed 22/12/2012).

[3] S. Sheu and C. Wu. Dynamic load balance algorithm (dlba) for ieee 802.11 wireless lan. *Tamkang Journal of Science and Engineering*, 1999. vol. 2, no. 1, pp.45-52.

[4] H. Velayos, V. Aleo, and G. Karlsson. Load balancing in overlapping wireless lan cells. *IEEE ICC 2004: Proceedings of International Conference on Communications*, June 2004. vol. 7, pp.3833-3836.

[5] A. J. Jardosh, K. Papagiannaki, E. M. Belding, K. C. Almeroth, G. Iannaccone, and B. Vinnakota. Green wlan: On-demand wlan infrastructures. *Springer, Journal on Mobile Networks and Applications*, December 2009. vol. 14, issue 6, pp.798-814.

[6] Y. Fukuda and Y. Oie. Decentralized access point selection architecture for wireless lans. *IEICE Transactions on Communications*, 2007. vol. E90-B, no. 9, pp.675-684.

[7] F. Guo and T-C. Chiueh. Scalable and robust wlan connectivity using access point array. In *IEEE DSN 2005: Proceedings of International Conference on Dependable Systems and Networks*, July 2005. pp.288-297.

[8] I. Jabri, N. Krommenacker, T. Divoux, and A. Soudani. Ieee 802.11 load balancing: An approach for qos enhancement. *Springer, International Journal on Wireless Information Networks*, 2008. vol. 15, pp.16-30.

[9] G. Bianchi. Performance evaluation and enhancement of the csma/ca mac protocol for 802.11 wireless lansanalysis of the ieee 802.11 distributed coordination function. *PIMRC'96: Seventh IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Oct 1996. vol. 2, pp.392-396.

[10] Q. Zhao, D. H. K. Tsang, and T. Sakurai. A simple model for nonsaturated ieee 802.11 dcf networks. *IEEE Communications Letters*, September 2009. vol. 12, no. 8, pp.563-565.

[11] IEEE Standard for Information technology Telecommunications and information exchange between systems-Local and metropolitan area networks Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, June 2007. IEEE Std 802.11-2007.

[12] H. Zhai, X. Chen, and Y. Fang. How well can the ieee 802.11 wireless lan suport quality of service. *IEEE Transaction on Wireless Communications*, November 2005. vol. 4, no. 6, pp.3084-3094.

[13] MadWifi. http://madwifi-project.org/ (last accessed 22/12/2012).

[14] I. Ramani and S. Savage. Syncscan: practical fast handoff for 802.11 infrastructure networks. *INFOCOM 2005: Proceeding of IEEE International Conference on Computer Communications*, 2005. vol. 1, pp.675-684.

[15] S. Tang, N. Taniguchi, O. Shagdar, M. Tamai, H. Yomo, A. Hasegawa, T. Ueda, R. Miura, and S. Obana. Potential throughput based access point selection. *APCC 2010: Proceeding of IEEE Asia-Pacific Conference on Communications*, 2010. vol. 1, pp.470- 475.

[16] N. Mishra, K. Chebroulu, B. Raman, and A. Pathak. Wake-on-wlan. *WWW 2006: Proceeding of ACM International Conference on World Wide Web*, 2006. pp.761- 769.

[17] IEEE 802.11v: IEEE Standard for Information Technology Telecommunications and information exchange between systems Local and metropolitan area networks Specific requirements, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: IEEE 802.11 Wireless Network Management, February 2011. IEEE Std 802.11v-2011.

[18] Scenargie network simulator. http://www.spacetime-eng.com (last accessed 22/12/2012).

Interference Aware Routing Using Localized Mobility Prediction for Multihomed Wireless Networks

Preetha Thulasiraman
Department of Electrical and Computer Engineering
Naval Postgraduate School
Monterey, CA, USA
pthulas1@nps.edu

Abstract— **In this paper, we present two novel algorithms to deal with mobility prediction and interference aware routing for multihomed wireless networks. First, a localized mobility prediction algorithm, LMP, is developed using the Hidden Markov Model (HMM) in which the multiple fixed relay nodes in the multihomed network architecture act as pseudo-base stations to locally maintain and deliver mobility information collected from surrounding mobile users. We show that the prediction accuracy of our proposed prediction algorithm is better than using common Markov chains to predict user location at a time instant t. We also show that our mobility prediction algorithm adapts better to a user node's change in movement. Second, we present a new interference aware routing algorithm in which the signal to interference noise ratio (SINR) is used as the routing metric to determine least interfering paths. The mobility prediction algorithm is used as input to the routing algorithm in order to accurately calculate the SINR value of a specific link at particular time instances. This information is used to perform route construction based on least interference. We solve the least interference routing problem using a minimal cost flow optimization framework. We show that the integration of the two algorithms outperforms conventional counterparts in the literature in terms of packet delivery ratio and end-to-end-delay. However, we also show that the tradeoff for increased network performance lies in the ability of the algorithm to scale to very large networks.**

Keywords – ***Interference; hidden markov model; SINR routing; prediction accuracy; minimum cost flow optimization***

I. Introduction

In recent years, services supported by mobile communications have expanded from simple voice traffic to various multimedia applications, resulting in the rise of 4G systems. These 4G cellular systems are required to provide high and homogeneous data rates over the complete cell coverage area while assuring a level of quality of service (QoS). In traditional cellular networks (in which each mobile station (MS) is directly connected to a base station (BS)), mobility management is performed by the base station. In such networks, mobility prediction is concerned with the user's path when it is within the coverage area of that base station. However, the traditional cellular architecture has a structural weakness in providing fair service because each user's QoS depends on its location and mobility within the cell. If a user is near the cell boundary, it experiences severe path loss and poor spectral efficiency compared to users near the base station. Thus, more resources need to be allocated for cell boundary users to obtain the same throughput.

Achieving the defined 4G objectives requires installing either a higher number of base stations, or integrating cellular and ad-hoc networking technologies. The integration of cellular and ad-hoc technologies, also referred to as Multihop Cellular Networks (MCN), has gained significant research attention given its capacity to achieve the 4G objectives [1], [2]. MCNs substitute the direct MS-BS link with multihop links using intermediate nodes (relays) to retransmit the information from source to destination. Various architectures are available to MCNs [3], including both fixed and mobile relays. In this paper we focus on MCNs with fixed relay nodes where the base station communicates directly with fixed relay nodes which in turn cooperatively relay information in an ad hoc fashion to other users in connectivity range. In this architecture, each fixed relay behaves as a "pseudo-base station" or "home" for the mobile users by providing services (i.e., routing and mobility management) that would normally be taken care of by the base station in a centralized manner. This is termed a *multihomed* MCN. The concept of multihoming has been extensively discussed in the context of Mobile IP [4] to improve network connectivity and manage mobility. Multihomed architectures have also been predominantly used to develop fault-tolerant routing protocols by ensuring that user nodes have multiple connection opportunities in the event that one home relay fails [5], [6].

A. Motivations and Related Work

Mobility management involving movement prediction relies on the availability of prior information on the user's mobility behavior. Recently, prediction schemes using variations of the Markov model, particularly the Hidden Markov Model (HMM) have been proposed for resource management purposes in ad hoc networks [7], [8]. These schemes use control theoretic frameworks to dynamically allocate resources to users. Similarly, mobility prediction in cellular networks has also been researched in [9], [10], [11].

The cooperation between fixed relays and the base station is the cornerstone for efficient communication at the network

layer. A mobile user is served by a nearby relay node that forwards packets (potentially over multiple wireless hops) to the base station. In addition to traffic forwarding and route decision making, the relays also have the responsibility of managing user mobility by collecting information regarding user movements from one home relay to another. This essentially reduces the burden on the base station by localizing mobility management.

A consequence of the increased use of fixed relays is the inherent interference that is induced. Wireless interference is influenced by node mobility and can lead to performance degradation. The time varying mobility patterns of the users (i.e., speed, direction etc.) can cause new interference to be induced at neighboring nodes [12]. Specifically, if a node n moves from an area of low interference, A, to one of high interference, B, then any transmission from n will contribute to the interference of area B.

Interference can be controlled/mitigated in the network layer i.e., with routing. In order to design an effective routing algorithm that mitigates the interference experiences of the wireless links, the mobility of the users must be considered. Mobility assisted routing has been studied in the literature for several years, more recently focusing on ad hoc and delay tolerant networks [13], [14]. However, none of these works discuss the direct impact of interference on the routing protocols. More recently, in [12], mobility aware routing using interference constraints was developed. However, the interference is modeled using the protocol model which induces binary conflicts (either two links interfere or they do not despite neighboring simultaneous transmissions) which is not true in practice. Our focus is on the use of the signal to interference noise ratio (SINR) interference model (also known as the physical interference model), which is based on practical transceiver designs of communication systems that treat interference as noise. Under the SINR model, a transmission is successful if and only if the SINR at the intended receiver exceeds a threshold such that the signal can be decoded with acceptable bit error probability. Although the SINR model has been shown to be more computationally complex than the protocol model, it also provides a more practical and realistic assessment of wireless interference [15]. Routing protocols using SINR to model interference have been studied in both static networks [16], [17], [18] and mobile networks [19]. However, although the work of [19] uses SINR for route selection, the mobility modeling is based on the random waypoint model, and therefore no specific mobility prediction is introduced. In addition, [19] does not correlate wireless interference with mobility.

Our objective in this paper is to study SINR and its relationship to interference based routing using localized mobility management information. We extend our work given in [1] by integrating an interference based routing structure into a refined mobility prediction algorithm.

B. Contributions and Organization

The contributions of this paper are two-fold. First, we propose a localized (distributed) mobility prediction (LMP) algorithm based on the HMM where the mobility information (i.e., location) of each user at a time instant t is collected by the corresponding home relay node for movement prediction purposes. Second, we develop a routing protocol which uses the location information of the mobile user to determine the interference level on links in its surrounding neighborhood. We use SINR as the routing metric to calculate the interference on a specific link. The SINR represents the link cost. We minimize the total cost of routing as a cost function of SINR while guaranteeing that the load on each link does not exceed its capacity, thereby determining least interfering paths from each user to the base station. The routing protocol and the proposed solution are solved using a combinatorial optimization technique, known as the minimum-cost flow problem in the operations research literature.

The rest of the paper is organized as follows: Section II describes the system model. In Section III, we discuss the LMP algorithm used in this paper while in Section IV the SINR based routing algorithm is developed. The performance evaluation of the LMP and SINR routing algorithms is discussed in Section V. We conclude the paper in Section VI.

II. System Model

The network topology used in this paper is based on the MCN model used in emerging 4G broadband wireless access networks [20]. The multihomed MCN that is the focus of this paper is shown in Fig. 1. As shown, the network architecture is based on three tiers of wireless devices: 1) user nodes which are the lowest tier; 2) relay nodes that route packets between the user and base station is the second tier; and 3) the base station is the highest tier. Each home relay interacts with a set of mobile users. It must be noted that a MS can directly interact with a BS rather than a home relay if it is closer to the BS than to the home relay. Let V_N denote the number of relay nodes and let V_M denote the number of users. The BS is connected to the wired infrastructure and behaves as a gateway to the Internet. The LMP algorithm that is used to predict the next location of each user node is handled by the individual home relays. Each home relay collects and maintains information regarding the movement of the mobile users connected to it.

To understand the interaction between the various components of our framework, we provide a block diagram given in Fig. 2. The block diagram shows the LMP algorithm and its relationship to the SINR based routing algorithm. The prediction of the user's movement is driven locally by a HMM that is performed by each home relay. This means that the HMM is used to represent the mobility pattern of the users. The current mobility information and the history of the user's past movements is used to make predictions. This information is maintained in the mobility database of each home relay which keeps track of users that are connected, were connected or will be connected (prediction) to the home relay. Specifically, the database keeps track of which users are connected to the relay and which users have moved away to another relay, base station or cell. The idea of the mobility database was originally developed in [7] and its implementation has been modified to suit the needs of the work presented in this paper.

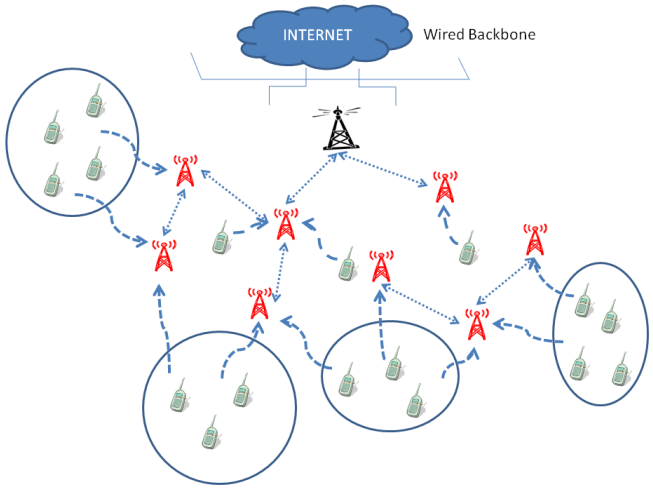


Fig. 1. Multihomed MCN where sets of user nodes are connected to a home relay and home relays communicate with other home relays in its transmission range to transmit information to the base station

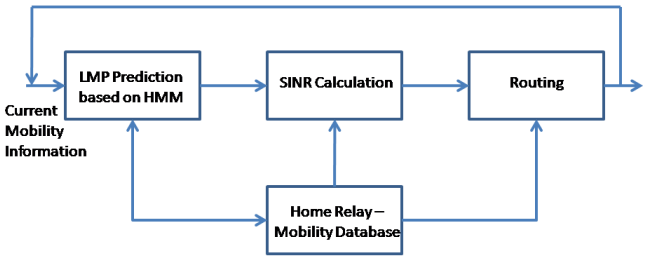


Fig. 2. Block diagram that illustrates the interaction between the LMP algorithm and the interference aware routing algorithm. The home relay runs the prediction algorithm and the SINR calculation for the routing procedure

The next predicted location of the mobile user, as determined by the home relay, is broadcast to other home relays so that they may update their databases accordingly. This updated information is then used to calculate the induced SINR interference at the receiver to proactively construct paths with least interference. The calculation of the SINR value at a time t in a mobile setting must be computed instantaneously. To facilitate the SINR calculation and the execution of the LMP and routing algorithms, it is assumed that the user nodes are quasi-mobile [21]; each user moves with a certain velocity and for a time T stays at one location before moving to a new random location.

III. Localized Mobility Prediction (LMP) Algorithm

The prediction problem discussed in this section aims to solve the following problem: *Consider a mobile user connected to relay node A. The user may move away from A to relay node B after some time. Using the history and transition paths, what is the likelihood that a user makes the transition from A to B?*

This problem has been dealt with using a Markov chain model [8]. However, the drawbacks of using a simple Markov chain model can be illustrated as follows. Referring to Fig. 3, consider a MCN with 4 relay nodes, A,B,C and D. Initially assume that a user connected to A moves from A to connect to any of the other relays, B,C or D. The transition from A to any of the other relay nodes may depend on proximity, signal strength, etc. The Markov model given in Fig. 3 shows the changes in direction as a sequence of probabilities based on past states. The transition probability for the next state is based on the most recent state. However, an external observer may not be able to see all of these transitions. Some transitions may be hidden from the observer by the user or the system. For instance, if a user connects to any of the relay nodes, the observer may only see the movement of the user from one relay to another but may not be able to determine which relay the user is connected to. Thus, the relay nodes are the hidden states and the locations are the observable states. Because there is no one-to-one mapping between these two states, the problem is to identify the relays corresponding to the location of the user.

A. Hidden Markov Model (HMM)

A HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. In a regular Markov model, the

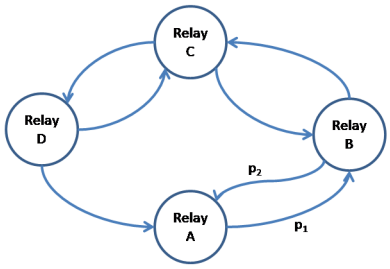


Fig. 3. Example to show a simple Markov chain that depicts the transitions of a mobile user to various relay nodes

state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a HMM, the state is not directly visible, but output, dependent on the state, is visible. A HMM has two kinds of stochastic variables: state variables (hidden variable) and the output variables (observable variable). A HMM can be defined as follows:

$O : \{o_1 o_2 ... o_N\}$ are the values of the observed sequences
$S : \{s_1 s_2 ... s_N\}$ are the N hidden states of the system
$\Pi : \{\pi\}$ is the initial state probabilities. π_i indicates the probability of starting in state i
$A = \{a_{ij}\}$ are the state transition probabilities where a_{ij} denotes the probability of moving from state i to j

$$a_{ij} = P(t_k = s_j | t_{k-1} = s_i)$$

$B = \{b_{ik}\}$ are the observation state probabilities where b_{ik} is the probability of emitting symbol k at state i

$$b_{ik} = P(o_k | t_k = s_j)$$

The 3-tuple (A, B, π) provides a complete specification of the HMM for the system considered in this paper.

To physically translate the HMM variables for the network at hand, O represents the relay node that a user is connected to presently. S represents which relay node a user will be connected to at a future time (where N denotes the number of relays) and Π is the set of state probabilities that indicate the likelihood that a user node is initially connected to a relay node i. A is the set of transition probabilities of a user node moving from a relay node i to a relay node j. Lastly, B represents the state probability of a user being connected to a relay node j given that the user started at relay node k. Essentially, B is the probability of an observed sequence. Given the parameters of the HMM model, the task is to compute the probability of an output sequence (i.e., which relay a user node is connected).

B. Localized Mobility Prediction Using HMM

To track the state of a mobile user we apply two approaches: 1) forward-backward algorithm and 2) re-estimation algorithm for the HMM parameters discussed above. The main steps of the tracking algorithm can be summarized as follows:

1) Apply HMM re-estimation algorithm to obtain initial estimates of (A, B, π) of the HMM.
2) Apply the HMM forward-backward estimation algorithm to predict at time t the next state of a user.
3) Obtain refined estimates of (A, B, π) by again applying the HMM re-estimation algorithm to the given observation sequences.

In mobile systems, up to date information regarding a user's movements is difficult to obtain. Estimation of the mobility model parameters must in general be made based on incomplete data. Due to physical constraints, transmission of location data may not take place frequently enough to allow precise tracking of the user's state at all times. The task of estimation from insufficient data involves two important aspects: (a) estimation and prediction of the user's movement behavior and (b) re-estimation of the model parameters based on incomplete information. These steps are performed at each home relay node during each observation time. We define the observation interval as the time during which observations (mobility information is collected) occur. The observation time is denoted as T, and is indexed by $1, 2, ..., T$. Time T is defined as the time during which the mobile user remains stationary. During this time, observations are collected for the LMP algorithm. Thus, the time during which the node remains stationary is the predicted state of the mobile network in the HMM.

1) Forward-Backward Algorithm: A forward-backward algorithm is an algorithm for computing the probability of a particular observation sequence in the context of hidden Markov models [22]. It is essentially an inference algorithm for HMM and consists of two steps. The first step of the algorithm computes a set of forward probabilities which provide the probability of observing the first k observations in the sequence and ending in each of the possible Markov model states (i.e., probability of ending up in any particular state given the first k observations). The second step of the algorithm computes a set of backward probabilities which provide the probability of observing the remaining observations given an initial state (i.e., probability of observing remaining observations given any starting point). These two sets of probabilities can then be combined to provide the probability of being in each state at a specific time during the observation sequence. The forward-backward algorithm can thus be used to find the most likely state for a hidden Markov model at any time.

For our model, we define the following forward and backward variables:

Forward variables represent the probability of an observation sequence $\{o_1 o_2 ... o_N\}$ and a state s_i at a time T. The forward variables, denoted as α, are determined as follows:

1) Initialization: $\alpha_i = \pi b_i(o_1)$, $1 \leq i \leq N$.
2) Induction: $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1})$, $1 \leq t \leq T-1$, $1 \leq j \leq N$.

Backward variables represent the probability of an observation sequence $\{o_1 o_2 ... o_N\}$ from $t+1$ to the end, given state s_i at time t. The backward variables, denoted as

β, are determined as follows:

1) Initialization: $\beta_T(i) = 1$, $1 \leq i \leq N$.
2) Induction: $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$ $1 \leq t \leq T-1$, $1 \leq j \leq N$.

The forward variables are computed inductively for $t = 1, 2, ..., T$. Similarly, the backward variables are computed inductively for $t = T, T-1, ..., 1$. After computing the forward and backward variables, a state estimate can be found. Let us define,

$$\gamma_t(n) = P[o_t; s_t = n]$$

as the probability that s is observed to be in state n at time t, where s is a user node. Then the estimate of s_t is given by

$$\hat{s}_t = \arg \max_{1 \leq n \leq N} \frac{\gamma_t(n)}{P[o_t]}, t = T, T-1, ..., 1$$

2) Re-estimation Algorithm: A simple iterative procedure for re-estimating the HMM parameters is reported in [22]. By applying the well-known EM (Expectation/Maximization) algorithm [23], it can be shown that this iterative procedure is increasing in likelihood. The overall computational complexity of the re-estimation algorithm is essentially proportional to T. Thus, the parameters of the HMM model can be estimated effectively within our framework.

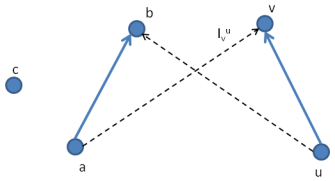
IV. SINR Based Routing Using Localized Mobility Prediction

This section will discuss the formulation of the SINR routing algorithm using the developed LMP algorithm.

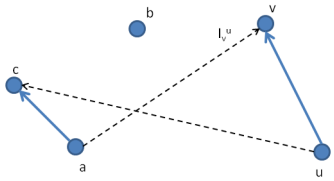
A. Challenge of Routing with Interference and Mobility

Using the LMP algorithm based on the HMM, we are able to track the movement of the users to determine which relay it is connected to. Given this information, routing from the connected relay to the base station can take place through multiple hops. Note that knowing to which relay a user is connected is imperative to the calculation of interference. To route in the presence of mobility and interference using link based metrics is a fundamental challenge. Under generic shortest path routing, the path length (which depends on the link metric) is the only factor that decides the best route between any source and the base station. Various examples of link metrics in the literature, namely Euclidean distance, depend solely on the two nodes forming the link. They are independent of the existence of other paths from other users and the BS or their shortest path routes. This, in turn, has led to the notion of link metrics and link-based routing. However, interference depends on the existence of other sources/intermediate relays and their spatial separation. Thus, the routing decision of a given source-base station pair becomes coupled to the routing decision of other source-BS pairs.

To illustrate this, assume node a is transmitting to next hop b and node u is transmitting to next hop v as shown in Fig. 4(a). According to the non-linear decay of power with distance, governed by $P_r(z) = P_t * z^{-\alpha}$ where P_t is the transmitted power, z is the distance between transmitter and receiver and



(a) Node a is transmitting to node b and node u is transmitting to node v



(b) Node a transmits to node c while node u continues to transmit to node v

Fig. 4. Illustration of the challenge of defining an interference aware routing metric in the presence of simultaneous transmissions and mobility

α is the pathloss exponent, the amount of interference at node v from transmitters other than u is given by $I_u^v = P_{ab} * z_{av}^{-\alpha}$. If node a was transmitting to a different node (i.e., node c), as shown in Fig. 4(b), then the amount of interference seen at node v would be different: $I_v^u = P_{ac} * z_{av}^{-\alpha}$. Note that P_{ab} is different from P_{ac}. Thus, the interference induced on link (u, v) (needed to compute its link metric) depends on the routing decision of transmitter a which, in turn, depends on the routing decision of transmitter u. Couple this scenario with mobility in which node a is moving, then a more refined time based routing metric is required to gauge both interference and the location of the node at that time.

To determine appropriate routing paths from the relay to the BS that are cognizant of interference, we use SINR as a routing metric. The SINR is an effective and practical metric to gauge link quality because it takes interference and noise as well as signal strength into account. Furthermore, with user nodes moving, poor links are unpredictable and thus SINR based routing decisions are useful to discover more robust paths.

B. Problem Formulation

For our analysis, we model the multihomed MCN as a graph, $G(V, E)$, where V is the set of nodes (relays, mobile users and base station inclusive) and E is the set of links. Let V_M be the set of mobile users and let V_N be the set of home relays, where $V_M, V_N \in V$. Note that the network has only one base station. The successful reception of a packet depends on the received signal strength, the interference caused by the simultaneously transmitting nodes, and the ambient noise level η. The SINR of a link (i, j) is given as follows

$$SINR_{ij} = \frac{P_j(i)}{\eta + \sum_{k \in V'} P_j(k)} \geq \beta \quad (1)$$

where $P_j(i)$ is the received power at node j due to node i, V' is the subset of nodes in the network that are transmitting simultaneously, and β is the SINR threshold. Our proposed routing protocol is implemented to route data using the least interfering path out of all path possibilities. If a link has a high SINR, it is an indication that it is experiencing low interference.

Each link (i, j) has an associated cost which is derived from the SINR value calculation. Each link also has an associated capacity denoted u_{ij}. The capacity is formulated using Shannon's formula, given in Eq.2.

$$u_{ij} = log_2(1 + SINR_{ij}) \quad (2)$$

In addition, the flow of packets from node i to its neighbor j over wireless link (i, j) is represented by f_{ij}.

C. SINR Based Routing

The position of each user node at time t affects the cumulative SINR on each link. The SINR is also affected by the path loss model and channel gain. The SINR at time t on link (i, j) is given by Eq.3,

$$SINR(t)_{ij} = \frac{G_{ij} P_j(i)(t)}{\eta + \sum_{k \in V'} G_{kj} P_j(k)(t)} \geq \beta \quad (3)$$

where G_{ij} is the channel gain on link (i, j) (in the simulations, the channel gain of each link is calculated using a Rayleigh fading model and an appropriate path loss factor), $P_j(i)(t)$ is the received power at node j due to node i at time t, and k is a simultaneously transmitting node. The corresponding capacity u_{ij} is then modified to be

$$u_{ij}(t) = log_2(1 + SINR_{ij}(t)) \quad (4)$$

The SINR is calculated during each observation time, $t \in T$.The cost of each link is associated with the SINR value obtained from Eq. 3.

In order to determine the least cost (least interfering) paths, we use the minimum cost flow optimization technique. In our case, the cost of a link is motivated by the amount of interference on that link due to neighboring transmissions and/or noise. As we are using SINR as the routing metric, the higher the SINR, the better the link quality. Therefore, we want to minimize the *inverse* of the SINR value.

The objective of the SINR routing problem is to deliver all the data packets generated by the user nodes to the base station in the most cost-effective (least interfering) manner without exceeding the link capacities. We can find least interfering paths for each user to the base station using the minimum cost (in this case minimum interference) flow optimization framework. Formally, the problem can be stated as follows.

$$\text{minimize} \quad \sum_{(i,j) \in E} SINR_{ij}(t)^{-1} f_{ij}(t) \quad (5)$$

subject to

$$\sum_{j:(i,j) \in E} f_{ij}(t) - \sum_{j:(j,i) \in E} f_{ji}(t) = d_i(t), \forall i \in V_M \quad (6)$$

$$\sum_{k:k \in V_N \cup BS} \Big(\sum_{j:(k,j) \in E} f_{kj}(t) - \sum_{j:(j,k) \in E} f_{jk}(t) \Big) = - \sum_{i:i \in V_M} d_i(t) \quad (7)$$

$$0 \leq f_{ij}(t) \leq u_{ij}(t) \quad (8)$$

$$f_{ij}(t) \in Z^+ \quad (9)$$

In the above formulation, d_i represents the rate at which the data packets are generated at user node i per unit time. The first constraint (Eq. 6) ensures flow conservation at each node. The second constraint (Eq. 7) ensures that the base station receives all the packets generated by all the nodes. The flow of packets on a link must not exceed its capacity and this is ensured by the third constraint (Eq. 8). The fourth constraint (Eq. 9) ensures that the (packet) flow values are integers.

The complexity of the above minimum cost flow problem is derived from [24] and shown to be $\bigcirc(\epsilon^{-2} E(E+V) log P)$ where E is the number of links in the network, V is number of nodes in the network (users plus relays) and P is an integer parameter that specifies the largest cost on the link (largest SINR value).

1) Solution: The above defined problem is similar to the minimum-cost flow problem, known in the operations research literature [25]. We will convert the above problem into the minimum-cost circulation problem as follows.

1) Add a super source x, and a super base station node y, to the graph $G(V, E)$.
2) Add directed links (x, i), connecting the super source x to node i, for all $i \in V_M \cup V_N$. Set costs of these links to 0 and the capacities to d_i.
3) Add directed links (j, y) connecting the base station and relay nodes to the super base station y. Set costs of these links to 0 and the capacities to infinity.
4) Add a directed link (y, x) connecting the super base station y to the super source x. Set the cost of the link (y, x) to $-|V|\beta$ and the capacity to infinity, where β is the minimum of any link cost (lower bound of SINR).
5) The modified graph is defined as $G'(V \cup \{x, y\}, E \cup E')$, where $E' = \{(x, i) : i \in V_N \cup V_M\} \cup \{(j, y) : j \in V_M \cup BS\} \cup \{(y, x)\}$.

The minimum-cost problem shown above is solved using the well-known minimum-cost flow algorithm given in [26]. An advantage of the minimum-cost flow algorithm is the integrality of flows. If all link capacities and expected data

rates of nodes are integers, then the minimum-cost flow algorithm can find paths with integral flow values.

2) Analysis of the Solution: The minimum path cost formulation given in Eqs. 5-9 determines the least interfering paths by minimizing the inverse of the SINR values of the links in the network. In addition, it also routes flows such that the link capacities are not violated.

Pushing more flow from x to y will decrease the overall cost of the flow due to the fact that the link from y back to x has sufficiently large negative cost. It is clear that the maximum flow is bounded from above by $F = d_1 + d_2 + ... + d_{|V_M|}$ because F is the maximum possible flow going out of x, the super source. There are two possibilities that have to be analyzed.

Case 1: $\sum_{i:i \in V_M} f_{xi} = \sum_{i:i \in V_M} d_i$

In this case, all the links of the form (x, i), $i \in V_M$ are saturated. The maximum-flow is restricted by the capacities of these links. Consider a link $(x, 1)$ having the capacity d_1. Since all the (x, i) links are saturated, the input flow at node 1 must be d_1 + $\sum_{j:(j,1) \in E} f_{j1}$ and the output flow must be equal to the input flow (flow conservation). There must be paths from node 1 to base stations which carry the flow d_1 + $\sum_{j:(j,1) \in E} f_{j1}$. The same argument holds for other nodes.

Case 2: $\sum_{i:i \in V_M} f_{xi} < \sum_{i:i \in V_M} d_i$

In this case, the maximum flow is restricted by the capacities on the actual links ($(i, j) \in E$) of the network. The minimum cost flow algorithm will identify the paths from the user node i to the base stations which carry the flow d'_i where $0 \leq d'_i \leq d_i$, $\forall i \in V_M$. The flow on the links (x, i) would be d'_i, $\forall i \in V_M$.

V. Performance Evaluation

A. Simulation Model and Performance Metrics

The LMP prediction algorithm and SINR based routing scheme have been simulated to verify their performance. The LMP prediction engine is first separately tested for accuracy in predicting the future mobility of users. For comparison, we use a generic Markov chain and a second-order Markov chain to gauge the prediction accuracy of the three methods. A second-order Markov chain can be defined as

$$P = P[Relay_{next} | Relay_{current}, Relay_{previous}]$$

When the users make first contact with a relay, there is no history of data from this user that can be utilized, so the initial parameters of the HMM are randomly generated using a uniform distribution (the number and locations of users and relays, relay-user associations and the initial transition probabilities are randomly generated). Once the users begin to move, its movement history is tracked and stored in the databases of each relay for prediction.

To evaluate the LMP algorithm, we look at its prediction accuracy. The prediction accuracy is one of the most important metrics for the verification of any mobility prediction algorithm. Prediction accuracy is defined as the ratio of the number of times a user moves to different relays to the ability of the system to predict the location. For example if node n moves to relay A and then to relay B, and our prediction algorithm predicts correctly that it moved to A but not B, then the prediction accuracy is 50%.

We use NS-2 to simulate our evaluations and use CPLEX to solve the optimization formulation for the minimum cost SINR based routing algorithm. The simulation environment is based on a 2250m x 2250m region with 14 relay nodes, 120 user nodes and one base station. The network environment is simulated using the NS-2 software platform, with the BS located at the center of the environment. The locations of the user nodes are randomly generated and then fixed in place. The propagation loss is modeled using the Rayleigh fading model. The Rayleigh fading model allows us to capture radio propagation signals that are not in the line of sight (i.e., when there are many objects in the environment that scatter the radio signal before it arrives at the receiver). The received power, $P_j(i)(t)$, is calculated according to the radio propagation model at the receiver. For simplicity, the transmission power of each relay node is set to 35dBm and the transmission power of each user is set to 24dBm. We also assume the radio transmission range to be 250m. The noise, η, is calculated as additive white Gaussian noise (AWGN) that is modeled as a Gaussian random variable. The pathloss exponent (LOS/NLOS) is set to 2.35/3.76. The threshold β for the SINR calculation is set to -18dB. The target SINR, for optimal network performance is -12dB. These values are defined specifically for voice data as is discussed in [27]. The standard deviation of the SINR is 0.5dB. With a data transmission rate of 2 Mbps, each run has been executed for 1000 seconds of simulation time. Constant bit rate (CBR) sources transmit UDP-based traffic at 4 packets per second and the data payload of each packet is 512 bytes long. The speed of the user nodes range from 1.5m/s to 5m/s. The simulated networks have 256 subcarriers with a system bandwidth of 2MHz. We also use different observation times, T. All results shown are an average of 20 different simulations.

To evaluate the SINR based routing scheme, we evaluate the following performance metrics:

- Packet Delivery Ratio: ratio of the number of data packets successfully delivered to the destination over the number of data packets sent by the source.
- End-to-End Delay: the average delay for a packet to reach from the source to the BS.
- Routing Overhead: Routing overhead is defined as the number of packet re-transmissions required because of packet drops/losses due to interference.

As benchmarks we compare with two interference aware routing metrics that use SINR as the routing metric, given in [16] and [19].

B. Simulation Results: Localized Mobility Prediction (LMP)

When the user nodes make first contact with a relay node, the initial, randomly generated parameters of the HMM are used. Each network that is simulated has 14 relay nodes (randomly placed), 120 user nodes (randomly placed) and 1 BS.

We first look at the performance of the LMP algorithm for two random users in the network and compare against the

Markov and 2nd-order Markov chains. Fig. 5 and Fig. 6 show the prediction accuracy in percentages for the two users in the network. From these figures we can conclude that the LMP has an advantage in prediction accuracy compared to the Markov and 2nd-order Markov chains. The results also show that the HMM can better adapt to a user's change in movement. In other words, the LMP learns faster than the generic Markov based approaches.

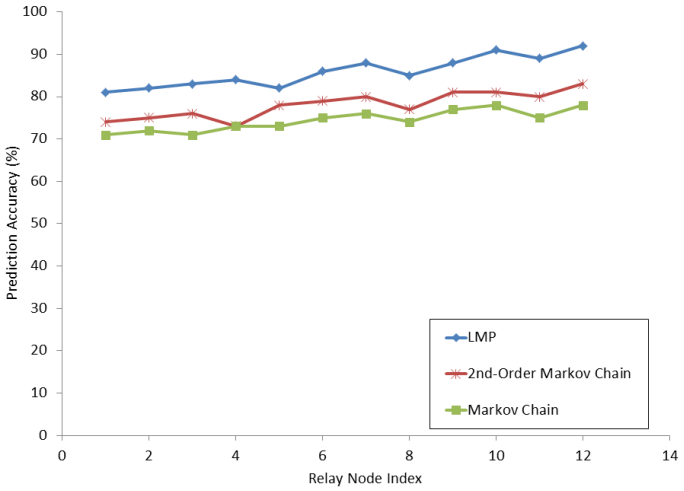


Fig. 5. Comparison of prediction accuracy for the proposed LMP algorithm, generic Markov chain and second-order Markov chain for User Node 1 in networks with 120 users, 14 relay nodes and 1 base station

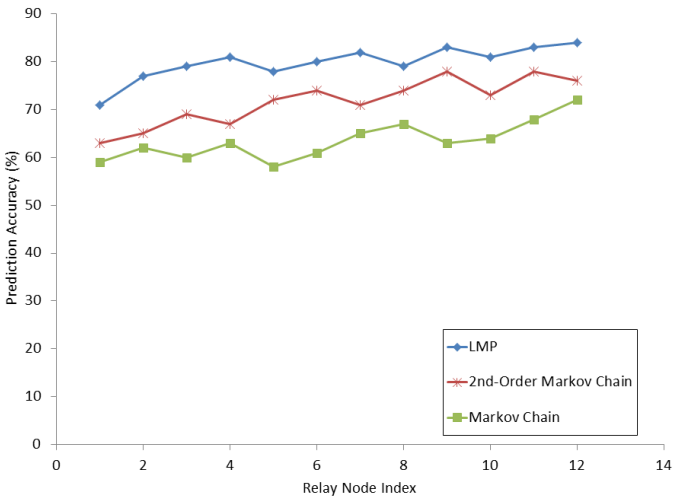


Fig. 6. Comparison of prediction accuracy for the proposed LMP algorithm, generic Markov chain and second-order Markov chain for User Node 2 in networks with 120 users, 14 relay nodes and 1 base station

C. Simulation Results of SINR Based Routing Algorithm

The performance of the SINR routing algorithm is evaluated compared to two SINR based routing algorithms given in [16] and [19]. In [16], an algorithm, 2-HEAR, is developed in which a routing metric is used such that a node calculates the SINR to its neighboring links based on a 2-hop interference range only. In [19], a modified version of the AODV routing algorithm is proposed in which SINR is used to calculate the route quality while using a random waypoint mobility model. We denote the above approaches as 2-HEAR and AODV-INT, respectively, in the simulation graphs. The same networks used in the LMP simulations of Section V-B are used in the simulations of the SINR routing algorithm. To calculate the SINR, we take the following steps.

We first evaluate the packet delivery ratio for our SINR based routing algorithm and its two relevant counterparts in the literature. In Fig. 7 and Fig. 8, the results of the packet delivery ratio for varying node speed and observation times ($T = 10$ms, $T = 1$ms, respectively) are shown. From the results it can be seen that our algorithm provides better packet delivery ratios when compared to the other approaches. We can justify the better performance of our results as follows: In 2-HEAR the SINR calculated by each node only includes those nodes within a 2-hop range which means that even if interference beyond this range occurs, it is not captured in the routing metric. If the interference level is high beyond the 2-hop range, packet drops may occur, requiring re-transmissions.The results of the algorithm from AODV-INT are better than 2-HEAR, however because AODV-INT does not use a specific mobility prediction model, it fails to capture precise interference information as is done in our proposed routing algorithm. It must be noted that the efficiency of the LMP-SINR routing algorithm is decreasing as speed increases (see Figs. 7 and 8). The faster the nodes move, the more likely the channels on which they are transmitting experience greater interference and fading. Thus, if the SINR is low, the efficacy of the LMP-SINR routing algorithm will decrease.

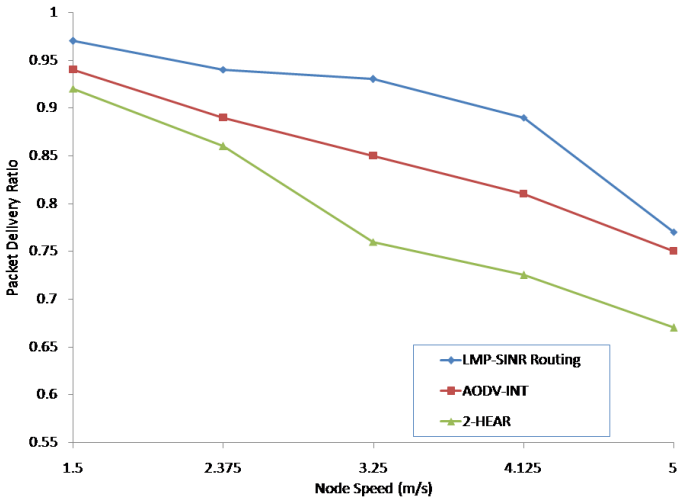


Fig. 7. Packet delivery ratio versus varying node speeds for $T = 10$ms

In addition, we also look at the effect of varying the observation time against the packet delivery ratio and show that with increasing T, the packet delivery ratio increases. The results are shown in Fig. 9, in which node speed is kept

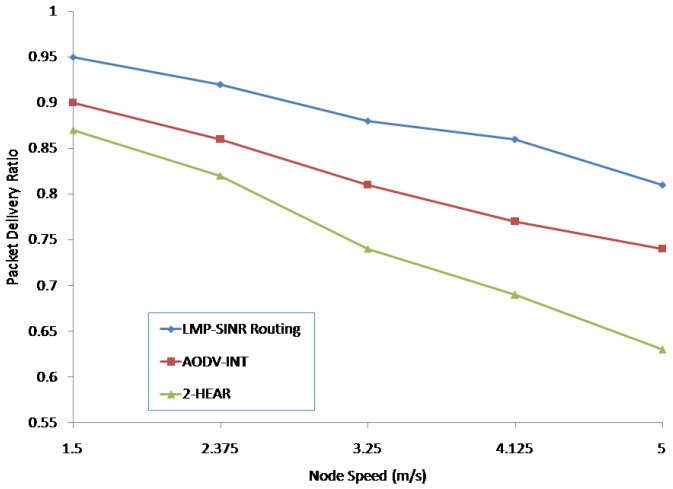


Fig. 8. Packet delivery ratio versus varying node speeds for $T = 1$ms

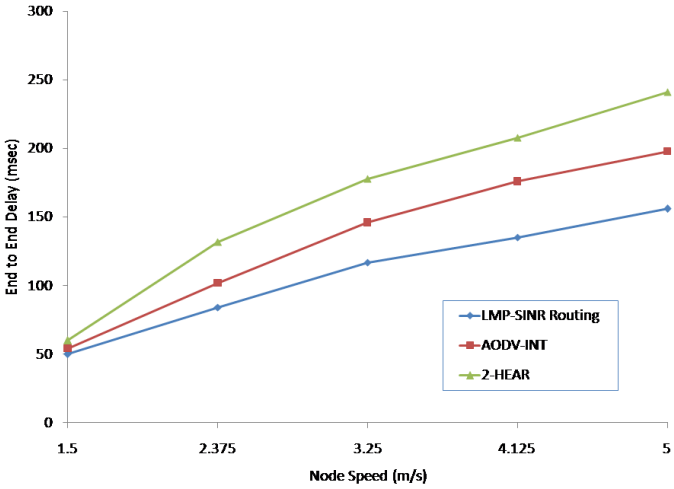


Fig. 10. End-to-end delay for $T = 1$ms and varying node speed

constant at 3m/s. This intuitively makes sense because T is essentially the amount of time used to observe the mobility of a node. The larger the value of T, the longer the LMP has to gather information leading to more accurate SINR calculation. This ultimately leads to better routes (less interference) and increases packet delivery ratios. This can also be seen in Figs. 7 and 8 in which packet delivery ratios are higher with $T = 10ms$.

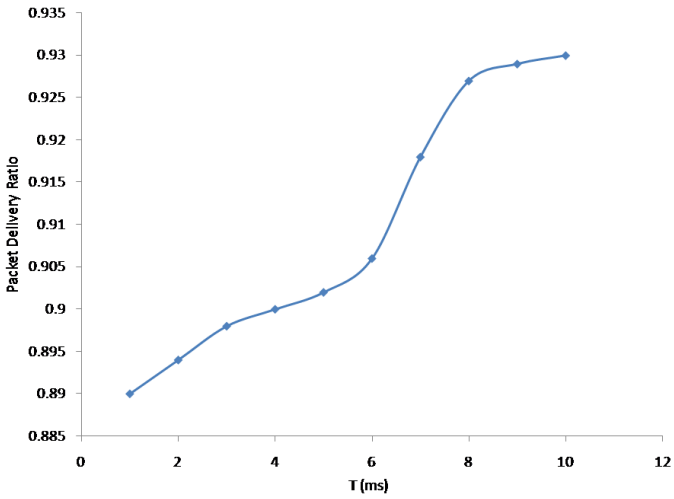


Fig. 9. Effect of varying T values on packet delivery ratio

We next evaluate the end-to-end delay of our algorithm for varying node speeds and $T = 1ms$. The results are shown in Fig. 10. The average end-to-end delay is improved compared to 2-HEAR and AODV-INT mainly due to more robust routes and less route discoveries. Note that the more reliable routes in our scheme significantly reduce the number of route discoveries and re-transmissions. This explanation also holds for the routing overhead produced by our proposed routing algorithm and that of 2-HEAR and AODV-INT. The routing overhead measured in this paper is that of how many packet re-transmissions are required when a routing path fails due to increased interference. The routing overhead is a measure of the number of data re-transmissions required per connection between a transmitter and receiver. Our calculation of interference is significantly more robust and inclusive than that of 2-HEAR and AODV-INT. Thus, the paths determined using our scheme are much more reliable, thereby indicating that the transmissions will be successful more often, requiring fewer re-transmissions of data. The results of the routing overhead, shown in Fig. 11, illustrate that the overhead of our scheme is less than that of the other two benchmarks.

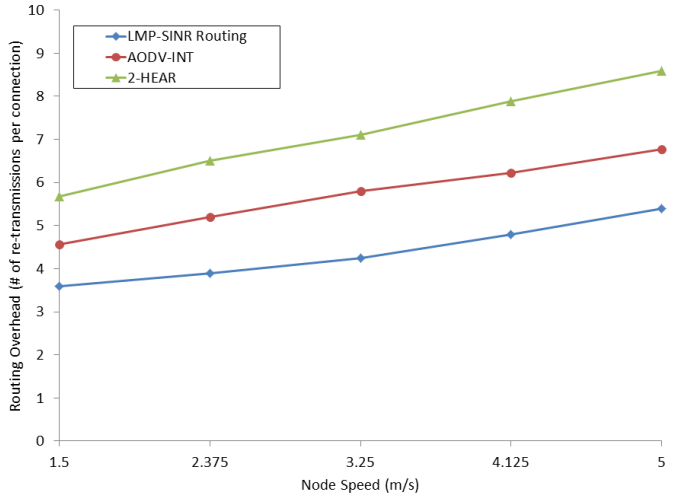


Fig. 11. Routing overhead for $T = 1$ms and varying node speed

Lastly, we look at the ability of our routing algorithm to scale to larger networks. The simulations shown in this paper were performed on networks with 120 user nodes and 14 relay nodes. When the algorithm is simulated on networks with 200 nodes or more, we found that the algorithm takes an

inordinate amount of time to converge. The primary reason for this is the time it takes to solve the minimum cost optimization formulation given in Eqs. 5-9. The running time of the optimization formulation, which is a function of the number of links and nodes in the network, does not scale to large networks. Thus, the performance improvement we see in terms of packet delivery ratio and end-to-end delay is a tradeoff for scalability.

VI. Conclusion

Mobility and wireless interference jointly influence the performance of wireless networks. In this paper we first developed a localized mobility prediction (LMP) algorithm using a Hidden Markov Model (HMM) for multihomed wireless networks. The mobility of each user is governed locally by individual home relays that capture and store mobility information. We then developed an interference aware routing algorithm using SINR as the routing metric, in which least interfering paths between each user and base station are found. In order to take into consideration the mobility of the user nodes within the routing algorithm, we use the LMP as input to the routing algorithm to predict the location of a user at time t. This predicted location is then used to proactively determine the SINR on each individual link. We formulated and solved the routing algorithm using a minimum cost (in our case minimum interference) flow optimization technique such that the link capacities are not violated. We showed that our LMP algorithm provides better prediction accuracy when compared to conventional Markov based mobility predictors. We also show that our SINR based routing algorithm guarantees minimum interference paths by increasing the packet delivery ratio and reducing latency compared to established SINR based routing approaches in the literature. In our future work, we plan to integrate the mobility of relay nodes to analyze the impact of SINR induced interference on routing and overall network performance.

Acknowledgement

This work was funded by the Research Initiation Program (RIP) at the Naval Postgraduate School, Monterey, CA, USA.

References

[1] P. Thulasiraman, "Mobility aware routing for multihomed wireless networks under interference constraints," in *Proceedings of International Conference on Emerging Network Intelligence (EMERGING)*, 2011, pp. 57–62.

[2] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: a new architecture for wireless communications," in *Proceedings of IEEE INFOCOM*, 2000, pp. 1273–1282.

[3] X.J. Li, B.-C. Seet, and P.H.J. Chong, "Multihop cellular networks: Technology and economics," *Computer Networks (Elsevier)*, vol. 52, no. 9, pp. 1825–1837, June 2008.

[4] Y. Li, D.-W. Kum, W.-K. Seo, and Y.-Z Cho, "A multihoming support scheme with localized shim protocol in proxy mobile ipv6," in *Proceedings of IEEE ICC*, 2009, pp. 1–5.

[5] P. Thulasiraman, S. Ramasubramanian, and M. Krunz, "Disjoint multipath routing to two distinct drains in a multi-drain sensor network," in *Proceedings of IEEE INFOCOM*, 2007, pp. 643–651.

[6] Y. Amir, C. Danilov, R. Musaloiu-Elefteri, and N. Rivera, "An interdomain routing protocol for multi-homed wireless mesh networks," in *Proceedings of IEEE WoWMoM*, 2007, pp. 1–10.

[7] P.S. Prasad and P. Agrawal, "Movement prediction in wireless networks using mobility traces," in *Proceedings of IEEE CCNC*, 2010, pp. 1–5.

[8] P.S. Prasad and P. Agrawal, "Mobility prediction for wireless network resource management," in *Proceedings of IEEE SSST*, 2009, pp. 98–102.

[9] W. Cui and X. Shen, "User movement tendency prediction and call admission control for cellular networks," in *Proceedings of IEEE ICC*, 2000, pp. 670–674.

[10] W.-S. Soh and H.S. Kim, "Dynamic bandwidth reservation in cellular networks using road topology based mobility prediction," in *Proceedings of IEEE INFOCOM*, 2004, pp. 2766–2777.

[11] H. Si, Y. Wang, J. Yuan, and X. Shan, "Mobility prediction in cellular network using hidden markov model," in *Proceedings of IEEE CCNC*, 2010, pp. 1–5.

[12] R. Langer, N. Bouabdallah, and R. Boutaba, "Mobility-aware clustering algorithms with interference constraints in wireless mesh networks," *Computer Networks*, vol. 53, no. 1, pp. 25–44, January 2009.

[13] L. Badia, N. Bui, M. Miozzo, M. Rossi, and M. Zorzi, "Mobility-aided routing in multi-hop heterogeneous networks with group mobility," in *Proceedings of IEEE GLOBECOM*, 2007, pp. 4915–4919.

[14] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Efficient routing in intermittently connected mobile networks: the multiple-copy case," *IEEE/ACM Transactions on Networking*, vol. 16, no. 1, pp. 77–90, February 2008.

[15] A. Iyer, C. Rosenberg, and A. Karnik, "What is the right model for wireless channel interference?," *IEEE Transactions on Wireless Communications*, vol. 8, no. 5, pp. 2662–2671, May 2009.

[16] R.M. Kortebi, Y. Gourhant, and N. Agoulmine, "On the use of sinr for interference-aware routing in wireless multi-hop networks," in *Proceedings of ACM MSWiM*, 2007, pp. 395–399.

[17] S. Kwon and N.B. Schroff, "Energy-efficient sinr-based routing for multihop wireless networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 5, May 2009.

[18] P. Thulasiraman, J. Chen, and X. Shen, "Multipath routing and max-min fair qos provisioning under interference constraints in wireless multihop networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 5, pp. 716–728, March 2011.

[19] J. Park, S. Moh, and I. Chung, "A multipath aodv routing protocol in mobile ad hoc networks with sinr-based route selection," in *Proceedings of IEEE International Symposium on Wireless Communication Systems (ISWCS)*, 2008, pp. 682–686.

[20] W. H. Park and S. Bahk, "Resource management policies for fixed relays in cellular networks," *Elsevier Computer Communications*, vol. 32, no. 34, pp. 703–711, March 2009.

[21] R.C. Ramos and L.F.G. Perez, "Quasi mobile ip-based architecture for seamless interworking between wlan and gprs networks," in *Proceedings of IEEE Conferences on Electrical and Electronics Engineering (CIE)*, 2005, pp. 455–458.

[22] S.-Z. Yu and H. Kobayashi, "Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden markov model," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1947–1951, May 2006.

[23] W. Turin, *Digital Transmission Systems*, McGraw Hill, 1998.

[24] L.K. Fleischer, "Approximating fractional multicommodity flow independent of the number of commodities," *SIAM Journal of Discrete Mathematics*, vol. 13, no. 4, pp. 505–520, October 2000.

[25] R. Ahuja, T. Magnanti, and J. Orlin, *Network Flows*, Prentice Hall, 1993.

[26] J. Orlin, "A faster strongly polynomial minimum cost flow algorithm," *Operations Research*, vol. 41, no. 2, pp. 338–350, 1993.

[27] M. El-Sayes and M.H. Ahmed, "An upper bound on sinr threshold for call admission control in multiple-class cdma systems with imperfect power-control," in *Proceddings of IEEE VTC-Spring*, 2007, pp. 2817–2821.

Performance Comparison of Enhanced Data Vortex Networks with Node Buffers and with Inter-cylinder Paths

Qimin Yang
Engineering Department
Harvey Mudd College
Claremont, USA
e-mail: qimin_yang@hmc.edu

***Abstract*— Optical switching fabric networks have become essential systems in high capacity communication and computing systems. This paper focuses on Data Vortex network architecture with two alternative implementations for improved performance. Either a buffer is added within the routing node or inter-cylinder paths are provided for enhanced routing performance. Since the extra hardware required for both implementations are the same, the network with better routing performance provides a better solution. A comparative study of the two methods is conducted with various load conditions and network redundancy. In addition to random traffic, performances under bursty traffic are also studied. The results have demonstrated that networks with inter-cylinder paths provide significantly lower latency and better throughput, and they are especially advantageous under bursty traffics. All results have shown that the approach with inter-cylinder paths provides more effective sharing of the routing resource within the network compared with the node buffering method. The difference in performance is also shown to be more dramatic under higher load conditions and for larger networks. Finally the comparison is also extended to a modified *4*-ary Data Vortex network, where traffic backpressure increasingly becomes a limiting factor due to deflection. Under medium to low redundant conditions, a similar performance trend is observed as that in regular binary Data Vortex network, where the inter-cylinder path method offer significant improvement in latency over the buffer node implementation, even though the latter also offers good improvement over the buffer-less *4*-ary network. A slight better performance in throughput is also shown in the inter-cylinder path method. In summary, we conclude that the inter-cylinder path enhancement provides a more attractive solution over the buffer based solution for various network operation conditions, especially promising for low redundant and high load conditions.**

Keywords- data vortex network; packet switched network; optical; network; buffering.

I. INTRODUCTION

Switching fabric networks are important subsystems in high capacity communication networks and computing systems. A typical space switch uses rich connectivity to handle dynamic traffic coming from a large number of input/output (I/O) ports while maintaining a high data throughput and small latencies. In high end multi-processor computing applications, the number of I/O ports or processors can reach thousands with each running at data rates of tens of Gbit/s. At the same time low latency (tens or hundreds of μs) must be maintained through such networks. Multistage self-routing network architectures often provide better system scalability, where each of the distributed routing nodes incorporates relatively simple routing logics. Such arrangement leads to cost-effective implementation and shorter delay due to simple processing at each stage. In order to provide higher data throughput, such networks can be implemented using optical fibre and optical switching technology.

Many recent researches have focused on developing optical switching fabric networks and network testbeds. In particular, this paper is a continuation of research presented in reference [1]. While it is relatively easy to achieve higher transmission bandwidth with Wavelength Division Multiplexing (WDM) within a single fibre, the routing logics and the handlings of traffic contention are hard to manage directly within the optical domain [2][3]. In particular, Data Vortex packet switched network architecture is developed for the ease of photonics implementation, and such networks are highly scalable to support a large number of I/O ports where each runs at high data rate and the network maintains a small routing latency [4]-[6]. The combination of its high spatial connectivity and an electronic traffic control mechanism among the routing nodes lead to bufferless operation and a much simpler routing logic within the nodes. Even though it uses deflection based routing, the spatial connectivity avoids large deflection penalty and reduces overall probability of deflection; therefore, it is advantageous compared with other commonly used interconnection architectures.

Previous researches on Data Vortex networks have focused on two main areas. One of the aspects has to do with physical implementation of the system. A small scale network testbed with 36 nodes and 12x12 I/O ports at Columbia University has been used to study various physical layer limitations. In particular as the number of node hops increases, optical signal to noise ratio (OSNR) and signal degradation were examined with various physical parameters. It has been shown that optical packets using an 8 wavelength payload at 10Gbit/s per channel can transverse 58 hops before a bit error rate (BER) of 10^{-9} is reached [7]. Therefore, the physical layer performance has shown promising scalability. Additional efforts are on switching device integration to support the size scalability. Current

testbed and system designs are based on semiconductor optical amplifier (SOA) switches because of their broad gain bandwidth and fast switching speed at nanoseconds, which is compatible with packet switching. Even though previous researches have not yet shown the same level as Data Vortex's potential sizes, several experimental works have demonstrated that a modular design can be used to build up a much larger matrix of SOA switches with required drivers and controls[8][9]. Integration related issues should be addressed for future study at much larger sizes and relevant cost scaling factor should also be explored in details. More recent researches on alternative switching devices based on silicon photonic technology can also provide potential solutions if these devices offer fast switching speeds while maintain low loss nature during the routing [10].

The second aspect focuses on enhancement in routing performance through network architecture designs. Although earlier researches have shown that with sufficient network redundancy, Data Vortex network scales to support a large number of I/O ports while achieving high throughput and low latency performance, at extremely high load conditions, and less redundant network conditions, the throughput tends to be limited by traffic backpressure in the deflection based routing. Therefore, network design researches may solve these issues with modified and enhanced functionality introduced in Data Vortex architecture. Simulation studies are typically conducted to examine the network performance under various traffic and operation conditions with different network sizes. There have been several approaches suggested to enhance the routing performance of the Data Vortex networks, especially for less ideal operating conditions [11]-[14]. In general, these performance enhancement methods require additional routing paths or routing resource, thus detailed cost and performance analysis must be carried out in comparison to the original network for a fair argument. There has been no comparison between different enhancement methods under the same operating condition, so this paper emphasizes such comparative study of two specific methods to contribute to the insights of the issues. The two methods, using node buffering and using extra inter-cylinder paths respectively, are of particular interests because they share the same cost with reasonable hardware increase in comparison to the original network. Among proposed, they are also relatively easy to implement thus more practical. The performance will be compared to each other as well as to the original Data Vortex networks. While random traffic is used for benchmark study, we also extend performance comparison under bursty traffics [15], which have not been previously studied within the enhanced networks. Simulation parameters are selected to focus on worse operation conditions such as low redundancy, high traffic load or bursty condition. In addition, recently a *k*-ary Data Vortex architecture based on multiple header bit processing at each stage has been proposed, which is shown to effectively reduce the latency when incorporated with buffer implementation. This is mainly due to smaller number of cylinders thus the forwarding delay is kept small in comparison to the overall delay [6]. Therefore, we also extend the proposed comparison study between two approaches in a *4*-ary Data Vortex network, and examine if the results for the original binary Data Vortex follow a similar trend in *4*-ary networks.

The paper is organized as follows: in Section II, the original Data Vortex network architecture is explained in details. In Section III, two previously proposed enhancement methods, the nodal buffering method as well as inter-cylinder path method are illustrated and compared in details. The routing performance comparison is provided in Section IV for various network conditions, and the comparison is extended to bursty traffic conditions as well as to *4*-ary Data Vortex networks. Finally the conclusion is given in Section V.

II. DATA VORTEX ARCHITECTURE

The Data Vortex architecture arranges its routing nodes in three dimensional multiple stage configuration as shown in Fig. 1. The size of the switching fabric is characterized by the height, H and angle, A of the cylinder. The number of cylinders is $C = \log_2 H + 1$ due to binary decoding routing process. The last cylinder is optional, but typically included to provide additional optical buffering for the output ports where electrical buffering is situated. Fig.1 shows routing path organization along each of the C=5 cylinders of the Data Vortex network with A=4, H=16. While the cylindrical levels (c=0 at the outermost cylinder to $c = \log_2 H$ at the innermost cylinder) provide the multiple levels in the routing stages, the angular dimension with repeated connection patterns provides multiple open paths to the destination therefore results in a much smaller latency penalty as deflection occurs. Inter-cylinder paths are not shown for a better view, and they are simply parallel links that maintain the height position of the packets when they propagate from outer to inner cylinders. These are used for forwarding purpose only between the different levels.

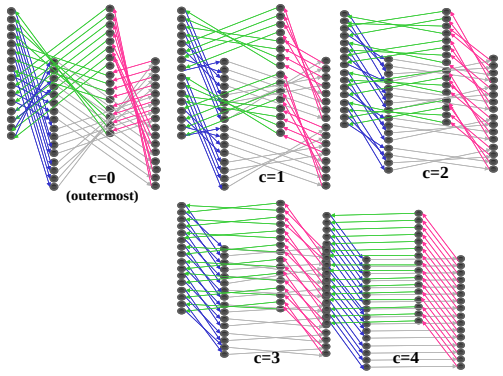


Figure 1. Data Vortex network with Angle=4, Height=16 and Cylinder=5 and its layout of routing node at different cylinders

A network in operation can connect I/O ports to active injection angle A_{in} with $1 \le A_{in} \le A$, and the ratio $\frac{A_{in}}{A}$ controls the network redundancy. For example, previous researches have shown that $\frac{A_{in}}{A} \cong \frac{1}{5}$ provides highly redundant condition, and it allows for each I/O port to reach above 95% injection rate even at full traffic load. This however requires an expensive implementation with $5 \times A_{in} \times C \times H$ number of routing nodes. Therefore, optimum choice of A_{in} should balance between the number of I/O ports and the desired routing performance.

Data Vortex networks operate in synchronous slotted fashion. Optical packets travel from the outermost cylinder to the innermost cylinder where the correct target height of the packet is located. To achieve simple self-routing process, each packet's destination height is encoded in binary. In the physical layer implementation, each of these binary bits is modulated onto a distinct wavelength, so that simple passive wavelength filtering can be used to extract and decode the single header bit h_i at the ith cylinder level. This is shown within the node structure in Fig. 2. Only a small amount of optical power is tapped and converted from optical to electronics (O/E) for header decoding purpose. Majority of the packet and power stays in optical domain as it travels through the network. Each node accepts either *West* (*W*) input (from the same cylinder) or *North* (*N*) input (from the outer cylinder or from the injection port). Only a single input can be present at the same time through traffic arbitration. The packet is routed either to *East* (*E*) (to the same cylinder) or to *South* (*S*) (to the inner cylinder) by turning on the proper SOA switch (SW). Each SOA provides power amplification to balance the power loss at the node due to tap and 3-dB power splitter between *E* and *S* paths, and its broad spectrum and fast nanosecond switching speed are appropriate for packet switching operation. The payload data is modulated using WDM technique as well, so that a typical packet of hundreds of nanoseconds could provide enough information per packet.

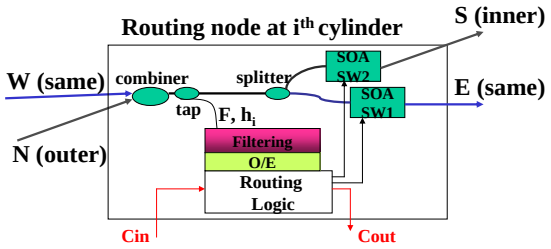


Figure 2. Routing node implementation

Data Vortex network combines a traffic control mechanism with deflection routing. Control signals stay in the electronic domain for simple implementation. As seen in the routing node in Fig. 2, a control signal C_{in} dictates whether *South* path to the inner cylinder is "open" or "blocking". Each routing node also generates a proper C_{out} to inform its outer cylinder neighbour node. The distributed control signal allows for the neighbouring nodes to coordinate properly and satisfy the single packet processing condition for each node. This can be illustrated in Fig. 3 in a triangle of routing nodes who shared the control signal path. Every time a packet is to stay at its current cylinder or to the *East* path, it creates a "blocking" control C_{out} for its outer cylinder contender. For example, if node A sends a packet to node B, it generates a "blocking" control for node C as shown in Fig.3. In the case the outer traffic receives a "blocking" control, the packet that is intended for *South* path will be deflected by staying on its current outer cylinder and wait for the next open path in two hops. In this example, packet of node C stays on cylinder c-1 until the next inter-cylinder path or corresponding control is open. The single packet routing arrangement eliminates optical buffers within the routing nodes as the network serves as virtual buffers as the packet travels on the cylinders.

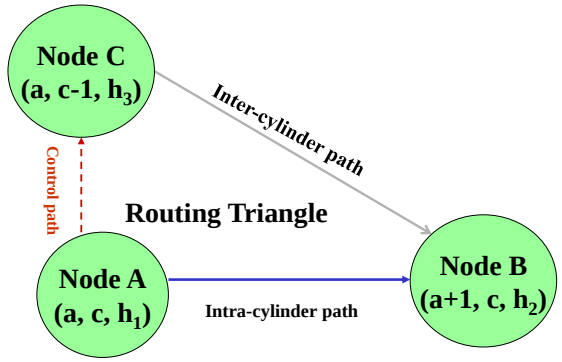


Figure 3. Control Signal in Routing Triangle

As mentioned, the last cylinder is typically added for optical buffering purpose so packets loop around the last cylinder at the same height position. Note that inter-cylinder paths and intra-cylinder paths are slightly different in length to allow for the establishment of the control signal and timing requirement. The inner cylinder nodes always make the routing decision slightly earlier than their outer neighbour to inform the traffic condition, so by making the inter-cylinder travel slightly shorter, packets arrive at the same node at the same time frame regardless of their origins. Detailed traffic control and routing performance have been reported in earlier studies [4]-[6]. Overall, Data Vortex networks maintain reasonable routing performance even as the networks scale up to thousands of I/O ports. In addition, many physical layer limitations have been studied and addressed in previous studies.

III. MODIFIED DATA VORTEX IMPLEMENTATION

As Data Vortex networks run at high load conditions or less redundant configurations, i.e., more input angles are attached to the I/O ports for incoming traffic, the traffic

backpressure could build up between the cylinders, so it takes longer to go through the network and the overall throughput also drops significantly. Due to the physical degradation of the optical signal through each node, reduction in latency is highly desired as well as maintaining the high data throughput. There have been several approaches suggested to enhance the routing performance of the Data Vortex networks with additional hardware. The detailed analysis of cost and performance comparison to the original network has been reported in earlier studies [11]-[13]. This paper emphasizes performance comparison of two methods using buffering and extra inter-cylinder paths respectively. Because the hardware increase in both methods is reasonably low and the costs are close to each other, a comparison of the two implementations under the same operation conditions is of great interests. In addition to previously reported random traffic performance, we have also extended the performance comparison for bursty traffic conditions. Section A provides an overview of the buffering method presented in [11], and section B provides an overview of the extra inter-cylinder path method presented in [12].

A. *Buffering*

The original Data Vortex network is attractive for its bufferless operation. However, for enhanced performance, separate buffers can be added within the routing nodes with slightly more complicated routing logic. This allows for less deflection when the packets wait in the buffer of the present node instead of circulating around the cylinders.

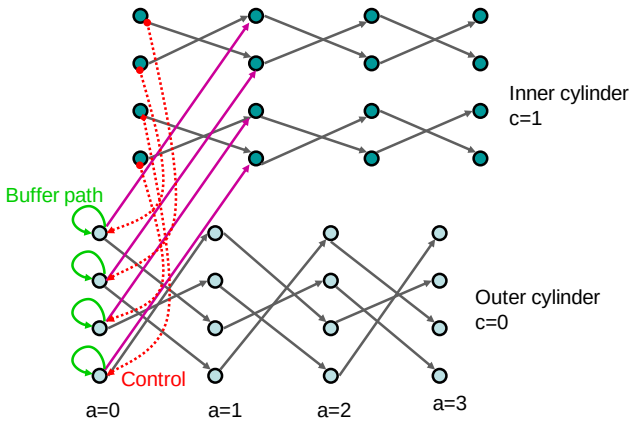


Figure 4. Data Vortex network with buffers within node shown at a=0.

Fig.4 shows the network implementation where nodes are arranged in the exact same fashion, except buffer paths are added within each node, as shown in an example for nodes at angle a=0. These buffer paths are simply delay lines with proper latency for routing purpose. The details of modified routing node are shown in Fig. 5. An additional switch (SW3) provides the third routing path to the buffer unit. Both the combiner and splitter will handle three potential inputs, so the splitting loss is slightly higher. The single packet routing principle is maintained so that only three SWs are required. In order to inform the presence of the traffic within the buffer path to maintain the single packet routing principle, the buffer unit must have at least two slot delays to allow for correct set up in timing of the control signal.

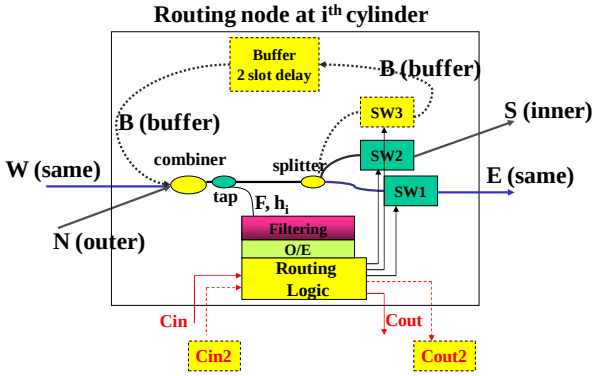


Figure 5. Routing node with buffer implementation: a 2-slot delay for buffer path is necessary to setup the control signal on time and additional controls C_{out2} are used to inform the state of buffer

Reference [11] also proposed a buffering scheme with a single slot delay, which is based on two simultaneous packets routing principle. While the routing performances are greatly improved, the required hardware is also significantly more because each node requires 6 SWs instead of 3 SWs. We are interested in a simpler and more cost-effective solution, so this study focuses on the buffer method shown in Fig.5 only that maintains a single packet routing principle through a two hop delay buffer. This implementation requires the network to have roughly 50% more hardware in number of switches and in routing paths compared to that in the original network. The modification of routing logic is minimal. If a packet is not able to reach *S* output, it will travel to the buffer unit and enter to the same node in two time slots. If the buffer packet is being processed, neither *W* nor *N* would accept inputs to maintain the single packet. As a result, priority is given to the packet within the buffer, and if there is no buffer traffic, then the same cylinder traffic gets the priority over the outer cylinder traffic as that in the original network. The additional control signal has to inform both the same cylinder neighbour and the outer cylinder neighbour to avoid contention.

B. *Inter cylinder paths*

In addition to buffering, there have been proposals for additional routing paths between the cylinders for enhanced routing performance [12][13]. The routing paths between the cylinders are critical resource and determine how fast traffic moves through the cylinders. Competition for these routing resource results in deflection thus builds up traffic backpressure. In this paper, we focus on the extra inter-cylinder path implementation as reported in [12], and a separate study has shown very similar enhancement results

for implementations in [12] and [13] under various traffic and network conditions. In the scheme shown in [12], we allow the packet to be routed to a secondary inter-cylinder path S_2 output if there is no other traffic (from regular *West* and *North* path) entering that same node. The addition the inter-cylinder path greatly improves the routing resource between cylinder levels. An additional injection path is also provided at each of the injection ports so that packets are less likely to be blocked by the traffic that is already circulating around the outermost cylinder. The setup of extra links and controls is shown in Fig. 6, and a detailed node implementation is shown in Fig. 7. The single packet routing rule is maintained for simplicity and an additional SOA switch (SOA-SW_3) is used to provide the third routing path as shown in the routing node. In this case, the additional control is necessary to inform the same cylinder traffic so that the traffic that goes to the regular S_1 output obtains the higher priority over the traffic that requires the S_2 output path. The secondary inter-cylinder path is of the same length as the original inter-cylinder path; therefore, it does not penalize packets that take the extra path in their delay. The implementation is merely trying to use the routing resource as much as possible while offer fairness to packets through the cylinders.

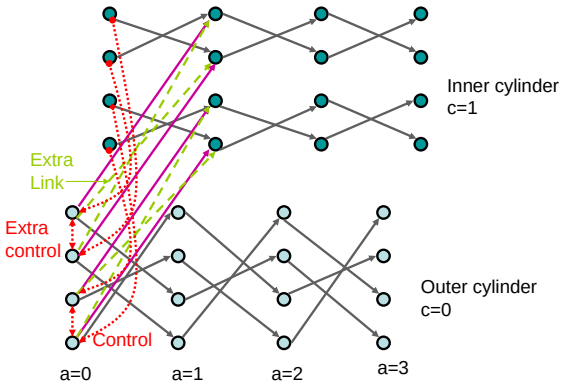


Figure 6. Extra inter cylinder path in Data Vortex network with required extra control

The height choice for the secondary inter-cylinder path must be such that the binary bits for all the previous cylinders maintain the same as those in the height of primary inter-cylinder path. As an example implementation, for a routing node at position of (a, c, h), its S_2 path connect to a node (a+1, c+1, h'), where h' can simply invert the $(c+1)^{th}$ bit of h where both height in binary format. Therefore, the first *c* header bits are locked the same to maintain the routing progress from the current node to either S path or S_2 path. The inter-cylinder path implementation requires about 50% more hardware in the number of switches and number of routing paths; therefore, it has comparable cost to the buffering implementation shown in section A.

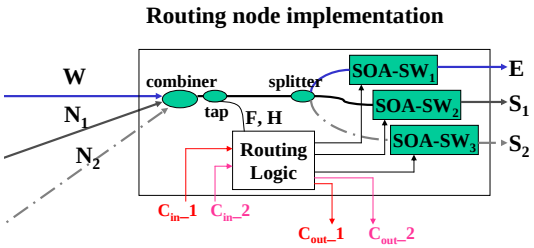


Figure 7. Modified routing node

IV. PERFORMANCE EVALUATION

In order to compare the effect of node buffering and inter-cylinder path for routing, a simulation in C/C++ is written to study the routing performance such as latency and data throughput. The compared networks are of the same size and same load conditions. The performance metric include average latency, latency distribution and network throughput. The average latency is measured for all the arrival packets for a long period of simulation time after the initial injection transient period. The network throughput is measured as the successful injection rate at the input port as previously reported. Latency distribution statistics are collected for arrival packets to see the range of the latency in packet switched operation. Once the packet reaches the correct target height, it exits the network immediately, therefore no angular resolution is considered in this simulation study. The performance evaluation extends beyond random traffic condition, and includes bursty traffic conditions as well as within a modified *k*-ary network implementation. These results are presented in section *A*, *B* and *C* respectively. The simulation runs sufficiently long for at least 5,000 clock cycles and the statistics are collected after steady state has been reached after the initial 500 clock cycles. All the results are presented with confidence level above 99% in comparison to a much longer simulation period or across various random seeds that are used to generate the traffic patterns. In all cases, the traffic load varies from 0.1 up to 1.0. Input angles A_{in} are typically chosen to be 3 or 5 to reflect medium to low redundant conditions in a network of *A*=5. Most simulations are carried out at a reasonably large size with H=256, and even higher sizes up to H=1024 are discussed for scalability study.

A. Performance comparison for random traffic

First random traffic pattern is studied to provide baseline performance. Random traffic indicates that each I/O port is independent, and they have a fixed probability of injecting packets, which depends on a set traffic load. Each packet slot also independently chooses its destination and its destination is uniformly distributed across all heights. Two enhancement methods are incorporated in a network of A=5, C=9 and H=256 as an example. Because both methods are for performance enhancement purpose when the Data Vortex network is heavily loaded or under less redundant operation, we choose the active injection angle to be A_{in}=3

and A_{in}=5 for the study. Keep in mind, for the buffer implementation, each buffer stay requires a two packet slots delay even though the number of node hop is one. The latency performance is measured in terms of packet slots to represent the physical delay.

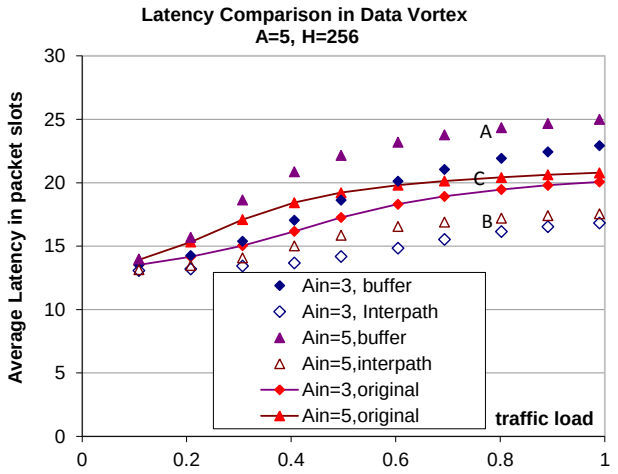


Figure 8. Latency comparison under various traffic load and redundant conditions

First, we examine the latency performance as shown in Fig. 8. For comparison purpose, the original network performances are shown as the solid lines. From these results, we can see that the inter-cylinder paths provide a smaller latency in general compared to that with an additional buffer within the routing node. In fact, the latency is worse for the case of node buffering compared to the original network especially at higher load conditions and less redundant network conditions. This is mainly because of the two hop delay requirement on the buffer path for timing requirement, which does not provide efficient reduction of latency even though the deflection events are reduced by keeping the packet at the open path to inner cylinder. The traffic backpressure remains significant because as the buffer packet re-enters the node for routing, there is no acceptance of additional traffic from neighbouring nodes. On the other hand, the inter-cylinder paths provide a better shared configuration of the redundant resource because when such resource is available, the additional routing paths always push more traffic through towards the inner cylinders. As a result, the traffic backpressure has been more effectively reduced. At the full load, the difference in latency in two methods is as large as 6 packet slots, which is 26.7% improvement if normalized.

The latency distribution is another important measure of the delay performance. In particular, we compare the latency distribution for A=5, H=256 with A_{in}=5 and at load of 0.8 for two implementation methods, i.e., network A and B shown in Fig. 8. The original network of the same condition or network C in Fig.8 is also shown as a reference. The latency distribution comparison is shown in Fig.9. A much narrower distribution is achieved in the inter-cylinder path approach, which dramatically reduces the average latency as previously shown in Fig.8.

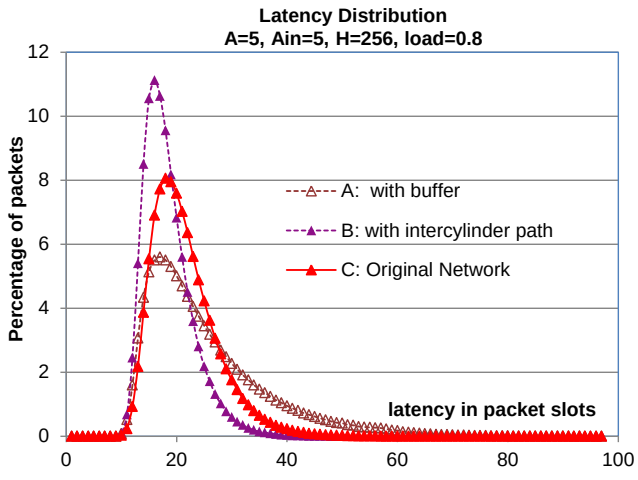


Figure 9. Latency distribution comparison for A, B and C in Figure 8.

The throughput performance comparison is shown in Fig. 10. A similar performance edge in inter-cylinder path implementation over buffer based implementation is reflected. In this rather busy network conditions, the buffer implementation has little improvement compared with the original networks, while the inter-cylinder path approach provides much more visible improvement. Both redundant conditions show very similar trend in comparison.

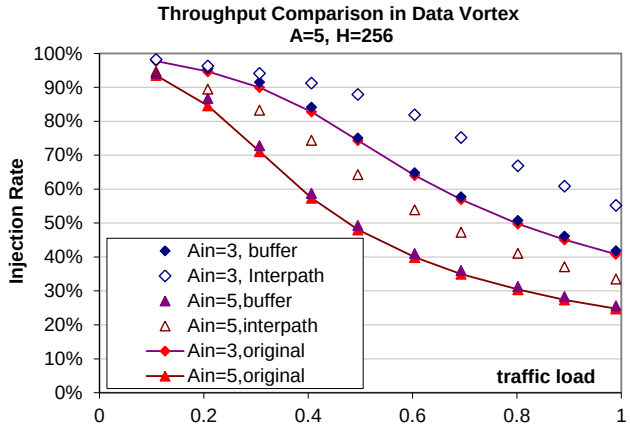


Figure 10. Throughput comparison under various load and redundant conditions

In reference [11], more detailed cost performance study is provided on this buffer implementation in comparison to the original network. Similar conclusion is provided that the overall the improvement in throughput and latency in this buffer scheme is rather limited and this implementation is only attractive for much lighter traffic conditions or more redundant networks. In our comparison for more heavily

loaded networks, the results have proved that the buffered implementation could even degrade the overall network performance once the system reaches saturation in load. On the other hand, the inter-path approach maintains the performance enhancement in both throughput and latency, and it provides a much more attractive solution for the same amount of hardware cost. Such performance enhancement clearly scales to very demanding network conditions as shown.

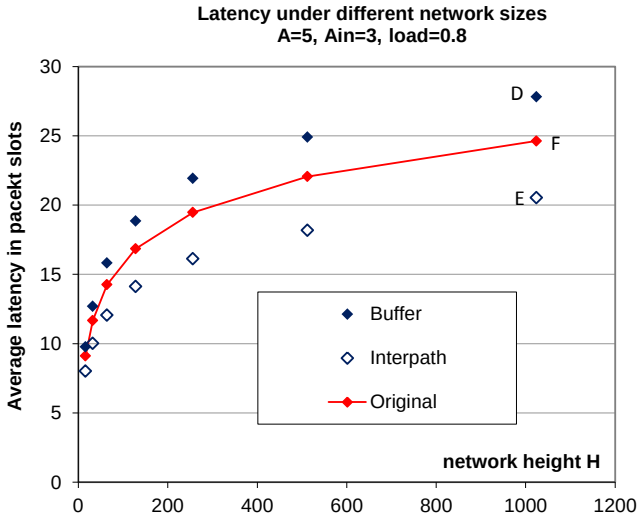


Figure 11. Latency performance comparison at different network sizes

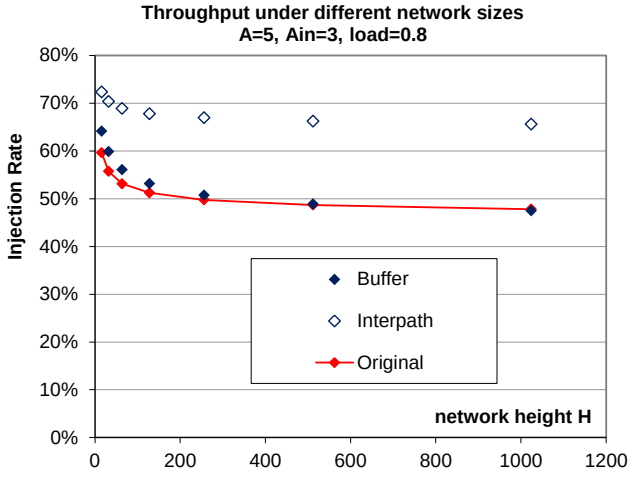


Figure 12. Throughput performance comparison at different network sizes

In order to study the scalability of such performance comparison, networks of different heights are also compared in the study. In Fig. 11 and Fig. 12, networks with both enhanced methods are compared with the original Data Vortex network with A=5 and A_{in}=3. All cases shown are with a medium to high traffic load of 0.8. It is shown that for all network sizes, the inter-path cylinder approach provides better performance over the buffer implementation, and there is especially significant difference for larger networks. In the case of H=1024, the latency difference between two methods is as large as 7 packet slots, which is 26.1% improvement if normalized. The throughput difference is as high as 18%, which is an improvement of 27.7% when normalized.

Finally, the latency distribution comparison for the two implementations for network height of H=1024, i.e., D and E shown in Fig.11 are also compared, and the original network F of the same condition is shown as a reference. As seen in Fig.13, the inter-cylinder path method provides much narrower latency distribution, and thus results in a much smaller average latency. As packets stay within the network less time on average, overall higher traffic throughput are achieved at the same time.

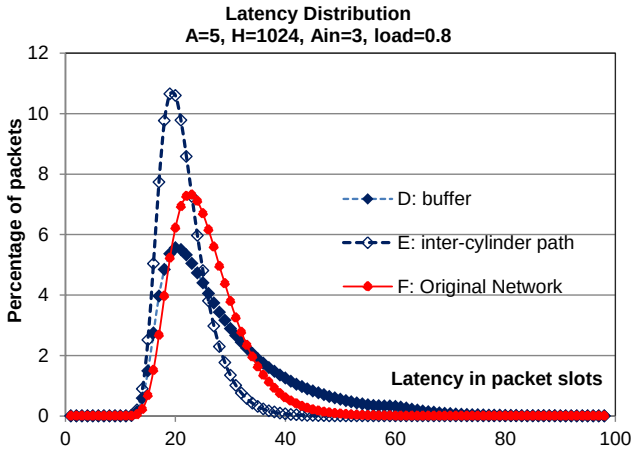


Figure 13. Latency distribution for D, E and F in Fig.11

B- Performance comparison for Bursty traffic

To show the performance comparison for even worse or more realistic traffic conditions, we extended the comparison to bursty traffic conditions. The effect of bursty traffic in the original Data Vortex has been studied thoroughly in reference [14][15], but the two enhanced networks with buffer and with inter-cylinder path have only been studied with random traffic. Here these enhancement networks will be subject to similar burstiness in traffic, and the results of modified architecture under the bursty traffic will be compared to each other, but also compared to random traffic performance.

As reported in [15], each ON period T_{on} and OFF period T_{off} are modeled by $T_{on} = \left\lfloor \frac{1}{U^{1/\alpha_{on}}} \right\rfloor$ and $T_{off} = \left\lfloor \frac{1}{U^{1/\alpha_{off}}} \right\rfloor$ respectively so that they follow rounded Pareto distributions. Here U is a random variable uniformly distributed over [0, 1], and $\lfloor \ \rfloor$ indicates the floor function. Parameters ($\alpha_{on,}\alpha_{off}$) specify the length of the consecutive injection slots and length of consecutive idle time slots, where consecutively injected packets are also of the same destination and treated as a burst. Each input port is modeled independently and traffic loads are averaged over different

input ports during the total simulation time. Table I indicates the burstiness parameter ($\alpha_{on,}\alpha_{off}$) and corresponding traffic load conditions used in the simulation study. In comparison to a random traffic of the same level of load, each burst goes to the same destination instead of individual slot; therefore such traffic pattern also causes hot spot in routing if the network is not properly designed.

Table I: Bursty parameter and actual load

Bursty Parameter α_{on}, α_{off}	Actual load
1.05, 8.0	0.856
1.05, 2.5	0.815
1.5, 5.0	0.712
1.5, 2.5	0.655
5.0, 5.0	0.5
5.0, 1.5	0.29

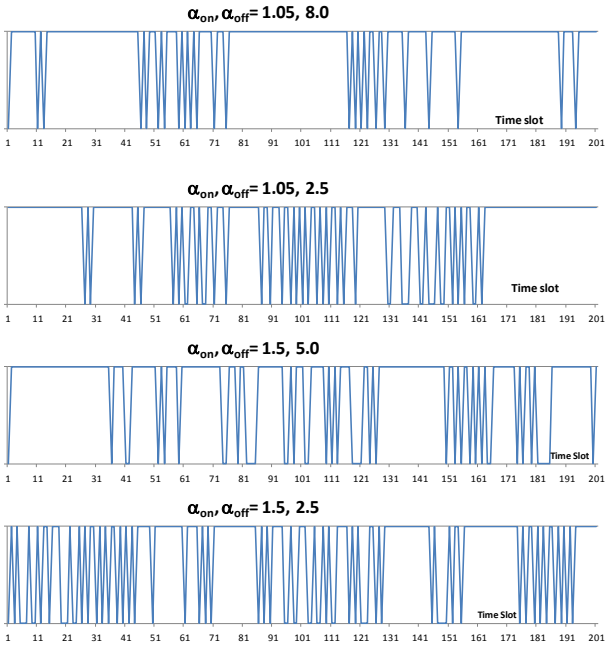


Figure 14. Bursty parameter and traffic patterns

Typical traffic patterns with the specified bursty parameters are shown in Figure 14 for comparison purpose. To really see the worst network condition, we present the comparison for the least redundant network condition with A_{in}=5.

Figure 15 and 16 show the latency and throughput performance respectively. As seen, the performance gain for inter-cylinder path implementation over buffer node implementation is even more obvious with bursty traffic conditions. In particular, the latency in buffer node networks shows a much worse uptrend (purples solid triangle) as the load increases for bursty traffic. The inter-cylinder path network on the other hand shows a very similar performance in latency between random and bursty traffic even at much bursty or higher load conditions. They almost follow the same range with much smaller sensitivity to the increases in load or burstiness. The throughput performance gain shows slight edge in inter-cylinder path method, but the performance difference is less obvious than the gain in latency performance.

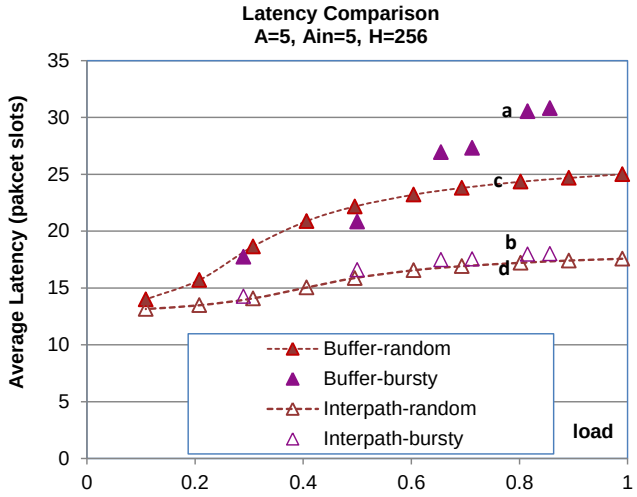


Figure 15. Latency performance comparison

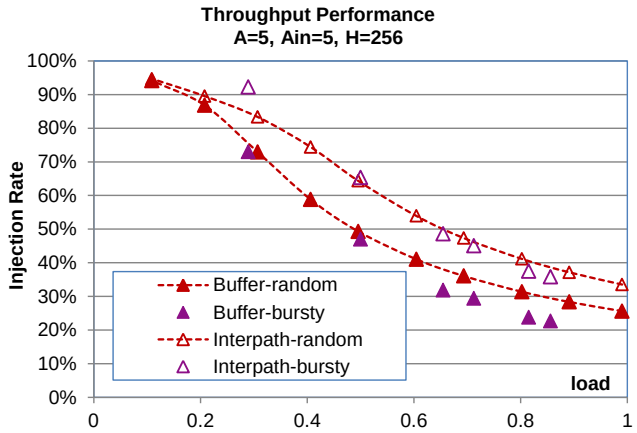


Figure 16. Throughput performance comparison

To further understand the latency performance, we also compare the latency distribution under various traffic conditions. In particular at load of 0.8 shown in Fig. 15, case *a* and *b* for bursty traffic and *c* and *d* for random traffic are compared and their latency distribution performance are represented in Fig. 17 and Fig. 18 respectively. As with the average delay, the distribution curve shows much narrower

range of packet latency with the inter-cylinder path implementation. On the other hand, with buffer implementation, the latency distribution shows much slower tail, and this is especially obvious in the case of bursty traffic conditions, which partially explained the much larger difference between case *a* and *b*, and this difference is more than difference between random traffic case *c* and *d* at a same level of traffic load.

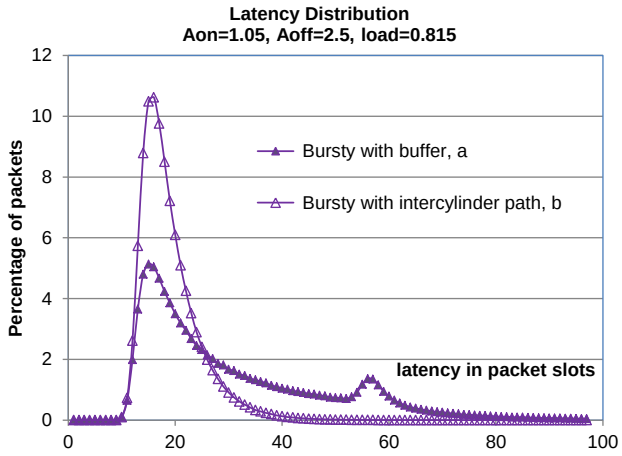


Figure 17. Latency distribution performance comparison

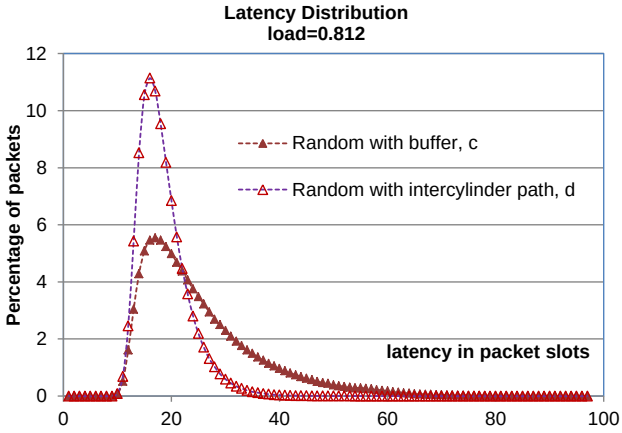


Figure 18. Latency distribution performance comparison

We also compared the performance difference for bursty traffic for different redundant conditions. As seen in Fig. 19 and 20, a similar trend is observed from a medium redundant network with A_{in}=3 in comparison to A_{in}=5 shown earlier. While the benefit is shown slightly less, it emphasizes the same conclusion that the inter-cylinder path implementation is more advantageous over buffer node implementation especially when the network is subject to worse traffic conditions or for load higher than 0.5.

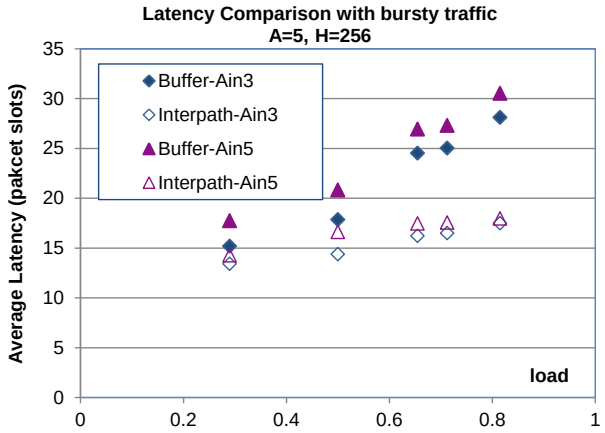


Figure 19. Latency performance comparison

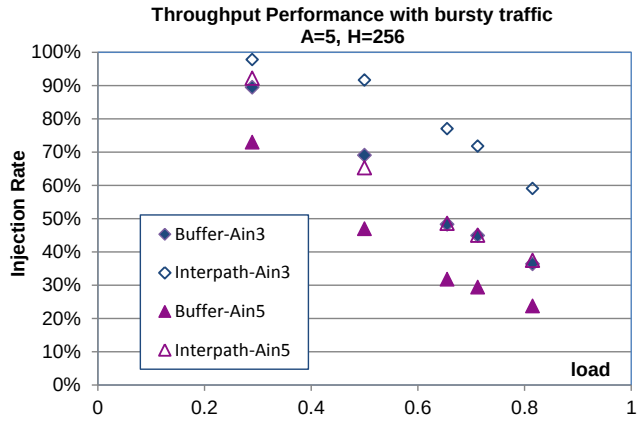


Figure 20. Throughput performance comparison

C- Performance comparison extended to 4-ary Data Vortex network

An alternative arrangement of *k*-ary Dava Vortex was proposed in a recent study, which is based on multiple header bit routing at each stage [16]. In particular, a *4*-ary network ($k = 4$) is shown to improve the latency performance due to the much smaller number of network cylinders and reduced forwarding latency. This is because number of cylinders is $C = \log_4 H + 1$ instead where each stage decodes two header bits ($\log_2 k = 2$) in a *4*-ary network. When incorporated with buffer implementation, such arrangement shows particular advantages because of smaller deflection penalty in comparison to bufferless *4*-ary network. Therefore, it is interesting to extend our comparison study between buffer implementation and inter-cylinder path implementation in the *4*-ary Data Vortex networks. Whether there is a same level of difference in two methods in their enhancement in *k*-ary network should be an interesting extension to the comparison results in the original binary Data Vortex.

As an example, a *4*-ary Data Vortex network is shown below in Fig. 21 which only requires $C = \log_4 H + 1 = 3$ cylinders for a network height of H=16. The routing node is modified as shown in Fig.22 so that the routing logic is based on two header bits and a similar traffic control mechanism is implemented to maintain the single packet processing principle. The routing path patterns of each cylinder can be constructed as shown [16].

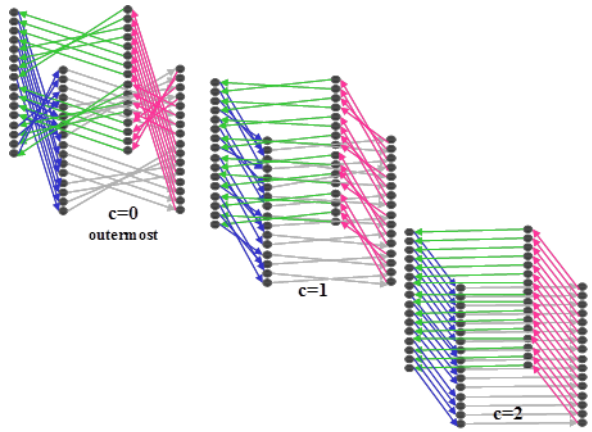


Figure 21. Routing patterns at each of the three cylinders in a *4*-ary decoding Data Vortex network. *A*=4, *H*=16, $C = \log_4 H + 1 = 3$

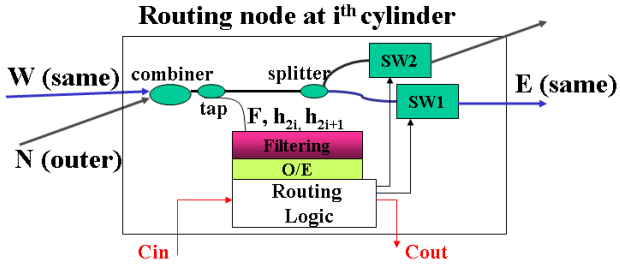


Figure 22. Routing node of *4*-ary network that requires two header bits detection

The network comparison is carried out for a same network height of H=256 as that in the binary network study. A=4 is chosen for the symmetry of the routing path patterns on the cylinder. Two redundant conditions with A_{in}=2 and A_{in}=4 are compared for the study to focus on medium to low redundant network conditions. We also include the original *4*-ary network without enhancement for reference, so the focus is on performance enhancement and comparison between two methods. Only random traffic is considered for this comparison.

The performance comparison in latency and throughput are shown in Fig. 23 and Fig. 24 respectively. A similar trend is observed in latency comparison, and under such redundant conditions, there is quite significant benefit of inter-cylinder path implementation over the buffer node implementation. For example, at full load condition, with least redundancy A_{in}=4, the difference in two methods in latency is as high as 6 packet slots, which is 28% if normalized. When compared to buffer-less 4-ary network, the gain in inter-cylinder path also reaches 19.4%. As seen, the effect of node buffering becomes limited, and it does not provide enhancement as in more relaxed traffic conditions [16]. The significant improvement in inter-cylinder path shows its effectiveness in routing. From the throughput performance, the difference is less significant, but still the inter-cylinder path provides slightly more improvement in comparison to the original network. Both buffer and inter-cylinder path offers better throughput than the buffer-less *4*-ary network, so traffic backpressure are reduced with both methods.

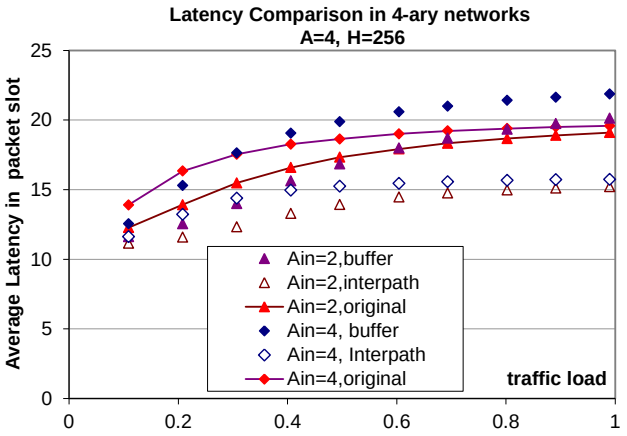


Figure 23. Latency performance comparison

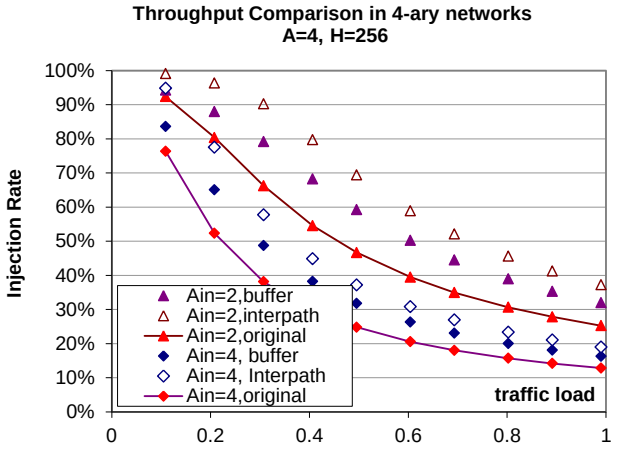


Figure 24. Throughput performance comparison

In summary, under medium to low redundant conditions, the *4*-ary Data Vortex networks performance follows a very similar trend as that in the binary networks. Overall, the inter-cylinder path implementation provides much more

significant improvement than the buffer node implementation. It is especially beneficial shown in the latency performance due to its single slot nature of the extra inter-cylinder path while the buffer is based on two slots delay. The throughput performance also shows slight edge for the inter-path cylinder method. The *4*-ary routing node implementation provides an overall reduction of the forwarding delay in comparison to binary network, but does not handle high traffic or less redundant conditions as well as binary which results lower throughput in general. The two enhancement methods provide greater benefits. Since the complexity and cost of implementation of two methods are the same, the inter-cylinder path offers a much more attractive solution because of its superior performance shown in all traffic conditions and network configurations.

V. CONCLUSIONS AND FUTURE WORKS

This study focuses on two different modification schemes for Data Vortex networks improvement. With similar hardware cost and complexity, the extra inter-cylinder paths provide a better configuration of the shared redundant routing resource. Such arrangement effectively reduces the traffic backpressure present in the original network at high load network conditions, and it provides much better performance in latency and data throughput than the modified network with buffering implementation. The extended study with bursty traffic further confirms the conclusion. The comparison in a *4*-ary Data Vortex network shows a similar trend, and the inter-cylinder path method offers obvious benefit over the buffer method, particularly in latency performance. Future developments in switching device integration are important and relevant for this investigation, and allow us to further quantify the benefits of different enhancement schemes. For future development in novel enhancement methods, researchers should consider not only the hardware cost but also the routing performance in both delay and throughput, especially for less ideal network operation conditions. Only a thorough study provides a fair and effective evaluation of the proposed solution.

REFERENCES

[1] Qimin Yang, "A Comparison Study on Data Vortex Packet Switched Networks with Redundant Buffers and with Inter-cylinder Paths", The third international conference on Emerging Network Intelligence (Emerging 2011), Lisbon, Portugal, November 20-25, 2011.

[2] Keren Bergman, Optical Fiber Telecommunications, B: Systems and Networks (Editor Ivan P. Kaminow, Tingye Li, Alan E. Willner), Chapter 19, "Optical interconnection networks in advanced computing systems", Academic Press.

[3] Ronald Luijten, Cyriel Minkenberg, Roe Hemenway, Michael Sauer, and Richard Grzybowski, "Viable opto-electronic HPC interconnect fabric", Proceedings of the 2005 ACM/IEEE SuperComputing, Seattle, pp. 18-18, November 2005.

[4] Roberto Gaudino, Guido A. Gavilanes Castilo, Fabio Neri, and Jorge M. Finochietto, "Can Simple Optical Switching Fabrics Scale to Terabit per Second Switch Capacities?", Journal of Optical Communication Networks, Vol.1, No.3, pp. B56-B68, August 2009.

[5] Odile Liboiron-Ladouceur, Assaf Shcham, Benjamin A. Small, Benjamin G.Lee, Howard Wang, Caroline P. Lai, Aleksandr Biberman, and Keren Bergman, " The Data Vortex Optical Packet Switched Interconnection Network", Journal of Lightwave Technology, Vol. 26, No. 13, pp. 1777-1789, July 2008.

[6] Cory Hawkins, Benjamin A. Small, D.Scott Wills, and Keren Bergman, "The Data Vortex, an All Optical Path Multicomputer Interconnection Network", IEEE Transactions on Parallel and Distributed Systems, Vol. 18, Issue 3, pp. 409-420, March 2007.

[7] Odile Liboiron-Ladouceur, Benjamin A. Small, and Keren Bergman, "Physical Layer Scalability of WDM Optical Packet Interconnection Networks", Journal of Lightwave Technolgy, Vol. 24, No. 1, pp. 262-270, January 2006.

[8] A.Wonfor, H.Wang, R.V.Penty, and I. H. White, "Large Port Count High-Speed Optical Switch Fabric for Use within Datacenters", Journal of Optical Communication Networks, Vol. 3, No. 8, pp. A32-39, August 2011.

[9] E.T. Aw, T. Lin, A. Wonfor, M. Glick, K.A. Williams, R.V. Penty, and I.H. White, "Layered Control to Enable Large Scale SOA Switch Fabric", 32nd European Conference and Exhibition on Optical Communication (ECOC 2006). Th1.2.5, Cannes, France, September 24-26, 2006.

[10] Xiaoliang Zhu, Qi Li, Johnnie Chan, Atiyah Ahsan, Hugo L. R. Lira, Michal Lipson, and Keren Bergman, "4 × 44 Gb/s Packet-Level Switching in a Second-Order Microring Switch", IEEE Photonics Technology Letters, Vol. 24, No. 17, pp. 1555-1557, September 2012.

[11] Assaf Shacham and Keren Bergman, "On contention resolution in the data vortex optical interconnection networks", Journal of Optical Networking, Vol.6, pp. 777-788, June 2007.

[12] Qimin Yang, "Enhanced control and routing paths in data vortex interconnection networks", Journal of Optical Networking, Vol. 6, No.12, pp. 1314-1322, December 2007.

[13] Neha Sharma, D. Chadha, and Vinod Chandra, "The augmented data vortex switch fabric: An all-optical packet switched interconnection network with enhanced fault tolerance", Optical Switching and Networking, Elsevier, Vol. 4**,** pp. 92-105, June 2007.

[14] Lianyong Dong, Qiang Dou, Quanyou Feng, and Wenhua Dou, " A Comparison Study of the Data Vortex Topologies with Different Parameter under Asymmetric I/O Mode", International Conference on Computer Science and Information Technology, pp.453-457, August 2008.

[15] Qimin Yang, Keren Bergman, "Performances of the Data Vortex Switch Architecture Under Nonuniform and Bursty Traffic", IEEE Journal of Lightwave Technology, Vol. 20, No.8, pp.1242-1247, August 2002.

[16] Qimin Yang, "Performance Evaluation of k-ary Data Vortex Networks with Bufferless and Buffered Routing Nodes", Asia Photonics and Communication Conference (ACP) 2009, pp. 1-2, Shanghai, China, November 2009.

Random Matrix Theory applied to the Estimation of Collision Multiplicities

Benoît Escrig
IRIT Laboratory
Université de Toulouse
Toulouse, France
E-mail: escrig@enseeiht.fr

***Abstract*—This paper presents two techniques in order to estimate the collision multiplicity, i.e., the number of users involved in a collision [1]. This estimation step is a key task in multi-packet reception approaches and in collision resolution techniques. The two techniques are proposed for IEEE 802.11 networks but they can be used in any OFDM-based system. The techniques are based on recent advances in random matrix theory and rely on eigenvalue statistics. Provided that the eigenvalues of the covariance matrix of the observations are above a given threshold, signal eigenvalues can be separated from noise eigenvalues since their respective probability density functions are converging toward two different laws: a Gaussian law for the signal eigenvalues and a Tracy-Widom law for the noise eigenvalues. The first technique has been designed for the white noise case, and the second technique has been designed for the colored noise case. The proposed techniques outperform current estimation techniques in terms of mean square error. Moreover, this paper reveals that, contrary to what is generally assumed in current multi-packet reception techniques, a single observation of the colliding signals is far from being sufficient to perform a reliable estimation of the collision multiplicities.**

***Index Terms*—multi-packet reception; collision multiplicity; model order selection; IEEE 802.11-based networks**

I. Introduction

Collisions are known to degrade the throughput of random access wireless networks, such as ad hoc networks[1]. A collision occurs when two or more user nodes access the channel simultaneously. Over the last decades, Medium Access Control (MAC) protocols have been designed with the rationale that all data packets from the colliding user nodes are lost when a collision occurs because the signals from all users mix.

It is is possible to retrieve part of the data packets that are involved in a collision, with approaches such as the capture effect [2]–[5]. Sophisticated techniques based on Multi-User Detection (MUD) allow the decoding of more than one data packets. MUD receivers have been successfully implemented in a wide range of application areas [6]. In this context, the number of colliding user is often needed to efficiently parameterize the MUD receivers.

Another approach is often invoked when a collision occurs. It consists in triggering a collision resolution (CR) mechanism where transmissions from the colliding users are re-scheduled in order to avoid another collision [2], [7]–[10]. In this context, CR mechanisms operate more efficiently when the number of colliding user is known. For instance, one could increase the contention window of a CR protocol with respect to the estimate of the number of colliding nodes. Note that our intention is not to propose a new CR mechanism but rather to provide a new parameter to improve the tuning of the mechanisms.

So, the purpose of this study is to estimate the number of colliding users, i.e., the collision multiplicity (CM) [1]. We focus here on the following scenario. The receiver at the destination node is processing a collision signal that consists of a mixture of signals from all users plus some Additive White Gaussian Noise (AWGN). The destination node is the node toward which all users involved in the collision are intending to send data [2]. So, from the observations of mixtures, the destination node has to estimate the number of original signals. This problem is a typical Model Order Selection (MOS) problem. MOS problems arise in the signal processing area and related areas such as signal array processing, radar, and sonar processing.

The MOS techniques are all based on the following rationale: the mixture of signals and noise can be decomposed into a noise subspace and a signal subspace, and the dimension of the signal subspace equals the number of signals. In order to perform this separation, the following steps are implemented. First, T observations of the mixture are gathered by the processing node. Then, the sample covariance matrix (SCM) of these observations is computed and an eigenvalue decomposition of this matrix is performed. These observations are obtained over T time-slots. The MOS techniques use the property that signal eigenvalues are much higher than noise eigenvalues, provided that the Signal to Noise Ratio (SNR) is high enough. When T eigenvalues are available and when K eigenvalues are significantly higher than the $T-K$ remaining eigenvalues, the conclusion is that there are K signals in the mixture.

The proposed approach has also been motivated by the following two observations: (i) many CM estimation techniques are based on the assumption that signal samples are

[1]In infrastructure-based networks, multiplexing techniques allow a fair bandwidth allocation among the users without any risk of collision.

[2]There are scenarios in which a collision occures between several source nodes transmitting toward several destination nodes. These scenarios are not addressed in this paper

uniformly distributed over a finite alphabet, i.e., signal samples are modulation symbols such as PSK or QAM symbols [11]–[13], and (ii) in many existing techniques, the number of observations is not much greater that the number of signals [7], [8].

Our point concerning (i) is to design estimation techniques that could be used in the context of Gaussian distributed samples. In this paper, we focus on the wireless standards that implement Orthogonal Frequency Division Multiplex (ODFM) transmissions; so signal samples can be considered as being Gaussian distributed and not uniformly distributed.[3] The issue that derives from (ii) is to clarify whether the number of observations can be made as small as $K + 1$ or $K + 2$ as in [7], [8] or not. Indeed, assuming that T and K are on the same order is in stark contrast with the typical assumptions that are used in signal processing [14]–[17].

We shall show that the proposed techniques outperform current techniques in terms of Mean Square Error (MSE) and that the number of observations T should be much higher than the number of signals K in order to obtain satisfactory MSE performance.

The rest of the paper is organized as follows. The introduction end ups with a Related Work subsection. The system model is introduced in Section II and some results on eigenvalue statistics are stated in Section III. The CMETs are described in Section IV. Simulation results are presented in Section V and a conclusion is drawn in the last section.

Related work

The first contributions, in the field of eigenvalue-based MOS techniques, were developed by Bartlett [18] and Lawley [19]. They propose a subjective setting of the threshold that is used to separate signal eigenvalues from noise eigenvalues. This approach is still used in some CR mechanisms [7], [8] in order to minimize the number of observations T. The algorithm starts with $T = 1$ and the number of observations is incremented by one each time the smallest eigenvalue is higher than a noise threshold. As soon as the smallest eigenvalue crosses the noise threshold at step T, that means that the T^{th} highest eigenvalue is a noise eigenvalue, and hence that there are $T - 1$ signals. Information theoretic criteria such as the Akaike Information Criterion (AIC) and the Minimum Description Length (MDL), developed by Wax and Kailath [17], have then been proposed in order to alleviate the limiting constraint imposed by the subjective setting of the separation threshold. The criteria are usually composed of a function that depends on the maximum likelihood estimator of the parameters of the model and a term that adjusts the first component to the context. This second term depends on the parameters of the system such as the number of samples per observation. The MDL have been used over several decades in several areas. An interesting review of this criterion can be found in [16].

[3]Note that signal samples in Code Division Multiple Access (CDMA) systems are also uniformly distributed since modulation symbols are multiplied by $+1/-1$ spreading codes.

The MDL is a consistent estimator of the number of signals when the number of observations, T, is fixed and when the number of samples per observations, denoted m, is such that $m \to +\infty$, provided that T is much larger than the number of signals K. Two limitations have recently been pointed out for this criterion.

First, the MDL have been shown to be inconsistent as the variance of the noise tends toward zero [15]. Second, the MDL is based on the distribution of the signal eigenvalues. It uses the property that sample eigenvalues, i.e., the eigenvalues that are obtained from the eigenvalue decomposition, are symmetrically distributed around the population eigenvalues, i.e., the theoretical eigenvalue. Basically, all the above-mentioned techniques are based on this property about signal eigenvalues. This centrality assumption makes sense in the T fixed, $m \to +\infty$ case. However, when T and m are on the same order, even when $T, m \to +\infty$, the previous property is no longer valid.

Recent advances in the Random Matrix Theory (RMT) field have brought into light several properties on the distribution of both signal and noise eigenvalues [20], [21]. These properties have been used in order to design new MOS techniques in the context where $T, m \to +\infty$ [16], [22]. These new techniques have been shown to outperform the classical MDL estimator.

Our purpose in this paper is to use these new properties in order to design new CM estimation techniques. Note that we are not in the context where $T, m \to +\infty$ but rather in the T fixed, $m \to +\infty$ case. The point is that current estimation techniques, whether they are based on the RMT or not, are based on the assumption that the number of observations T is much larger than the number of signals K. In our context, we want to minimize T with respect to K, so we are dealing with a context where T is in the order of K. Since the new RMT-based MOS techniques are performing in a wider range of parameter values, we believe that they can be considered as relevant candidates for our objective.

II. System Model

In this paper, we take in interest collision resolution algorithms and, more precisely, the estimation of the number of stations that are involved in a collision. We consider the scenario where K stations are simultaneously transmitting data packets to the same destination node (see Fig. 1).

We assume that the destination code and the colliding stations are all equipped with a single antenna. This is a worst case scenario. When the destination node is equipped with several antennas, observations are gathered more rapidly. The K colliding stations are transmitting OFDM frames of m samples each. The OFDM signal samples are Gaussian distributed with zero mean and variance unity. Moreover, the OFDM signals from the colliding users are assumed to be uncorrelated. This assumption makes sense since each transmission is affected by both a different Doppler shift and a different Doppler spread. The destination node receives T observations. So it is assumed that the destination can trigger transmissions from the colliding nodes. Note that T must be

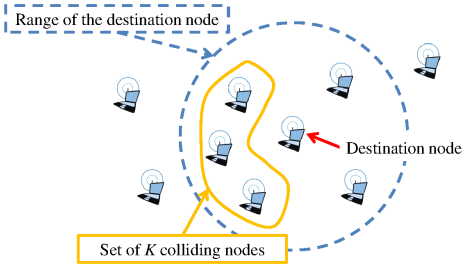


Fig. 1. Collision Scenario with $K = 3$ colliding nodes.

larger than K to allow an efficient estimation of K. This will be discussed in the simulation section. The destination triggers transmissions from the colliding users by sending a feedback frame. This frame serves also as a synchronization frame for user transmissions. Hence, the K users can be assumed to be coarsely synchronized in time. So this system model is similar to a source separation problem when K signals are impinging on a T sensor array. The $T \times 1$ observation vectors $\mathbf{y}_i$ can be written as

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}_i, \qquad i = 0, 1, \ldots, m \tag{1}$$

where $\mathbf{x}_i \sim \mathbb{C}\mathcal{N}_K(\mathbf{0}, \mathbf{R_x})$ are $K \times 1$ complex Gaussian vectors with zero mean and covariance matrix $\mathbf{R_x}$, $\mathbf{b}_i \sim \mathbb{C}\mathcal{N}_T(\mathbf{0}, \mathbf{R_b})$ are $T \times 1$ complex noise vectors that are Gaussian distributed with zero mean and noise covariance matrix $\mathbf{R_b}$. In the case of a white noise, we have that $\mathbf{R_b} = \sigma^2\mathbf{I}_T$ where σ^2 denotes the noise variance and $\mathbf{I}_T$ is the $T \times T$ identity matrix. The channel matrix $\mathbf{A}$ is considered as an unknown $T \times K$ non-random matrix. For our study, we assume that the coefficients of $\mathbf{A}$ are modeled as circularly symmetric Gaussian coefficients with power unity (Rayleigh fading). The channels gains are assumed to have constant values over the duration of the frame and change randomly from one frame to another. This corresponds to the typical block-fading channel assumption.

The observations $\mathbf{y}_i$ can be whitened by the following transformation

$$\mathbf{y}_i^\dagger = \mathbf{R_b}^{-1/2}\mathbf{y}_i$$

provided that the noise covariance matrix $\mathbf{R_b}$ is known *a priori* and is nonsingular, where $\mathbf{R_b}^{+1/2}$ is the Hermitian nonnegative definite square root of $\mathbf{R_b}$. The transformation simply reduces to a normalization step in the case of a white Gaussian noise. The covariance matrix $\mathbf{R}$ of the snapshots $\mathbf{y}_i$ is given by

$$\mathbf{R} = \mathbf{A}\mathbf{R_x}\mathbf{A}^{\mathrm{H}} + \mathbf{R_b} = \Psi + \mathbf{R_b}$$

with $^{\mathrm{H}}$ denoting the complex conjugate, the signal and noise vectors being independent. We assume that the channel matrix $\mathbf{A}$ is full rank and that the signal covariance matrix $\mathbf{R_x}$ is nonsingular so that the rank of Ψ is $\min(K, T)$. Hence, when $T \leq K$, there are T non-zero eigenvalues in the matrix Ψ and when $T > K$, there are K non-zero eigenvalues. This property is used in [7], [8] where the number of observations, T, is incremented until $\mathbf{R}$ has zero eigenvalues. When the whitening transformation is applied, the covariance matrix $\mathbf{R}^\dagger$ of the whitened observations, denoted $\mathbf{R}^\dagger$, is defined as

$$\mathbf{R}^\dagger = \mathbf{R_b}^{-1/2}\mathbf{R}\mathbf{R_b}^{-1/2} = \mathbf{R_b}^{-1/2}\Psi\mathbf{R_b}^{-1/2} + \mathbf{I}_T$$

The population eigenvalues $\mathbf{R}^\dagger$, denoted $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_T$, are such that

$$\lambda_i > 1 \quad \text{for } 1 \leq i \leq \min(K, T) \tag{2}$$
$$\lambda_i = 1 \quad \text{for } \min(K, T) < i \leq T \tag{3}$$

In most approaches, T is much larger that K so the conditions in (2) and (3) should be written as $1 \leq i \leq K$ and $K < i \leq T$ respectively. However, in some CR algorithms such as the ones in [7] and in [8], (2) and (3) are used in this way, since they are initialized with $T = 1$ and T is incremented by one each time the T^{th} smallest eigenvalue is not detected has being a noise eigenvalue. In any case, the estimation of K can be performed from the multiplicity of the λ_i equalling one. When $\mathbf{R}$ and $\mathbf{R_b}$ are known, and when the rank of $\mathbf{R_b}^{-1}\Psi$ is K, the CM estimation can be easily performed from the multiplicity of the λ_i equalling one. Otherwise, when $\mathbf{R}$ and $\mathbf{R_b}$ are unknown and have to be estimated, the sample covariance matrix (SCM) of the observations $\mathbf{y}_i$, denoted $\widehat{\mathbf{R}}$, must be computed

$$\widehat{\mathbf{R}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_i\mathbf{y}_i^{\mathrm{H}}$$

The SCM of the noise, denoted $\widehat{\mathbf{R}}_\mathbf{b}$, must be also computed using

$$\widehat{\mathbf{R}}_\mathbf{b} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{b}_j\mathbf{b}_j^{\mathrm{H}}$$

Note that the $\mathbf{b}_j$, $1 \leq j \leq N$ are independent noise-only samples. We assume that we can get noise-only samples by using idle time slots. In that case, the estimation of K is based on the eigenvalue decomposition of the SCM $\widehat{\mathbf{R}}^\dagger$

$$\widehat{\mathbf{R}}^\dagger = \frac{1}{m}\sum_{i=1}^{m}\mathbf{y}_i^\dagger(\mathbf{y}_i^\dagger)^{\mathrm{H}} = \widehat{\mathbf{R}}_\mathbf{b}^{-1/2}\widehat{\mathbf{R}}\widehat{\mathbf{R}}_\mathbf{b}^{-1/2} \tag{4}$$

The sample eigenvalues of the SCM matrix $\widehat{\mathbf{R}}^\dagger$ are denoted $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_T$.

III. Some Results on Eigenvalue Statistics

In the following, we investigate new results in RMT. In particular, we provide new characterizations of the eigenvalues of SCMs. We first begin with the signal-free case and then address the signal bearing case.

A. Signal-free Case

We start with the signal-free case where no user is transmitting, so $K = 0$. As a consequence, we have that

$$\begin{aligned} \mathbf{R} &= \mathbf{R_b} \\ \mathbf{R}^{\dagger} &= \mathbf{I}_T \end{aligned} \tag{5}$$

and all population eigenvalues λ_i are all equal to one. From the system model and (5), the vectors $\mathbf{y}_i$ are T-dimensional complex Gaussian independent vectors such that $\mathbf{y}_i \sim \mathbb{C}\mathcal{N}_T(\mathbf{0}, \mathbf{R_b})$.

From Property 6.1, we get that $m\widehat{\mathbf{R}}$ is a T-variate complex Wishart matrix with m degrees of freedom and covariance matrix $\mathbf{R_b}$, i.e.,

$$m\widehat{\mathbf{R}} \sim \mathbb{C}W_T(m, \mathbf{R_b})$$

Similarly, since we have that $\mathbf{y}_i^{\dagger} \sim \mathbb{C}\mathcal{N}_T(\mathbf{0}, \mathbf{I}_T)$, we have that

$$m\widehat{\mathbf{R}}^{\dagger} \sim \mathbb{C}W_T(m, \mathbf{I}_T)$$

From Property 6.2, we now characterize the eigenvalues of $\widehat{\mathbf{R}}$ in the signal-free case and in the presence of a white Gaussian noise [16], [20], [23], [24].

Corollary 3.1: In the signal-free case, when $\mathbf{y}_i \sim \mathbb{C}\mathcal{N}_T(\mathbf{0}, \lambda\mathbf{I}_T)$, the Empirical Spectral Distribution (ESD) of the $T \times T$ random matrix $\widehat{\mathbf{R}}^{\dagger}$ converges almost surely to the Marčenko-Pastur law in (14) (see Fig. 2).

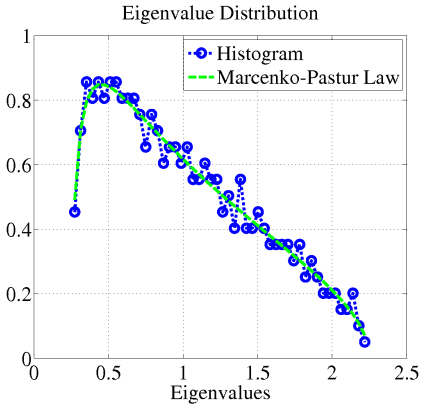


Fig. 2. Distribution of the eigenvalues of a matrix $\widehat{\mathbf{R}}$ in the signal-free case with $T = 500$ and $m = 2000$ ($\lambda = 1$).

When the snapshots $\mathbf{y}_i$ are whitened and become $\mathbf{y}_i^{\dagger}$, a new distribution for the ESD must be considered [22].

From Property 6.3, we now characterize the eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$ in the signal-free case in the presence of an arbitrarily colored Gaussian noise [20], [22]–[24].

Corollary 3.2: In the signal-free case, when $\mathbf{y}_i^{\dagger} \sim \mathbb{C}\mathcal{N}_T(\mathbf{0}, \mathbf{I}_T)$, the ESD of the $T \times T$ random matrix $\widehat{\mathbf{R}}^{\dagger}$ converges almost surely to the modified Marčenko-Pastur law in (16) (see Fig. 3).

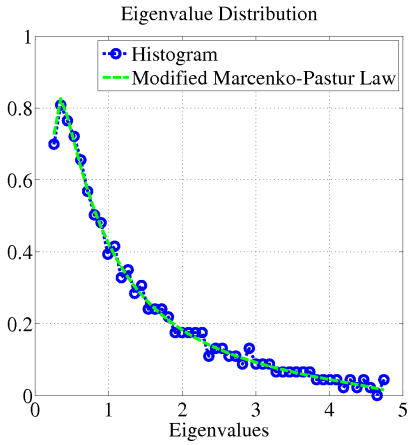


Fig. 3. Distribution of the eigenvalues of a matrix $\widehat{\mathbf{R}}^{\dagger}$ in the signal-free case with $T = 500$ and $m = 2000$.

In the signal-free case, we are interested in the eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$ or, equivalently, in the eigenvalues θ that satisfy

$$\widehat{\mathbf{R}}v = \theta\widehat{\mathbf{R}}_{\mathbf{b}}v \tag{6}$$

We can rewrite (6) as follows

$$m\widehat{\mathbf{R}}v = (\frac{m}{N}\theta)N\widehat{\mathbf{R}}_{\mathbf{b}}v \tag{7}$$

Since $m\widehat{\mathbf{R}} \sim \mathbb{C}W_T(m, \mathbf{R_b})$ and $N\widehat{\mathbf{R}}_{\mathbf{b}} \sim \mathbb{C}W_T(N, \mathbf{R_b})$, (7) is similar to (20). So, using Property 6.5, the largest eigenvalue $\hat{\lambda}_1$ that satisfies (7) is Tracay-Widom (TW) distributed and we can derive the following property.

Property 3.1: The largest eigenvalue $\hat{\lambda}_1$ that satisfies (6) is TW distributed, i.e.,

$$\mathrm{P}\left\{\frac{\log(\frac{m}{N}\hat{\lambda}_1) - \mu(T,m,N)}{\sigma(T,m,N)} \leq x\right\} \rightarrow TW_{\mathbb{C}}(x)$$

The pdf of the largest eigenvalue $\hat{\lambda}_1$ is represented in Fig. 4. Note that this characterization uses explicitly the double Wishart setting that has been motivated by the need to whiten the observations when the additive Gaussian noise of the channel is not white. Another and simpler characterization of $\hat{\lambda}_1$ has also be proposed in [20], [21].

Property 3.2: In the signal-free case, the whitened snapshots $\mathbf{y}_i^{\dagger}$ are $\mathcal{N}_T(\mathbf{0}, \mathbf{I}_T)$ and the largest eigenvalue $\hat{\lambda}_1$ of the SCM $\widehat{\mathbf{R}}^{\dagger}$ is Tracy-Widom distributed. When $T, m \rightarrow \infty$ such that $T/m \rightarrow c \in (0, \infty)$,

$$\mathrm{P}\left[\frac{m\hat{\lambda}_1 - \mu_{T,m}}{\sigma_{T,m}} \leq x\right] \rightarrow TW_{\mathbb{C}}(x)$$

where

$$\begin{aligned} \mu_{T,m} &= (\sqrt{T} + \sqrt{m})^2 \\ \sigma_{T,m} &= (\sqrt{T} + \sqrt{m})(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{m}})^{1/3} \end{aligned}$$

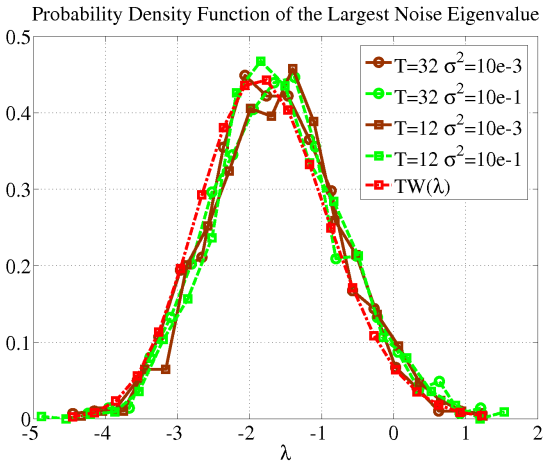


Fig. 4. Probability density function of the largest eigenvalue of a matrix $\widehat{\mathbf{R}}^{\dagger}$ in the signal-free case with $m = 61440$ and different values for T and the noise variance σ^2.

When the snapshots $\mathbf{y}_i^{\dagger}$ are $\mathcal{N}_T(\mathbf{0}, \sigma^2\mathbf{I}_T)$, the convergence limit of $m\hat{\lambda}_1$ is $\sigma^2(\sqrt{T} + \sqrt{m})^2$. This corresponds to the non-normalized case. Note that the convergence rate to the $TW_{\mathbb{C}}(x)$ distribution function is $\mathcal{O}(T^{-1/3})$. When the parameters m and T are not so large, which is practically the case when we want to reduce the number of observations, the convergence rate to the TW distribution is rather $\mathcal{O}(T^{-2/3})$ provided that the mean and standard deviation have been modified appropriately .

B. Signal Bearing Case

When there are K signals and when $T \to \infty$, the limiting ESD of $\widehat{\mathbf{R}}^{\dagger}$ still converges to a Marčenko-Pastur distribution. This is because all eigenvalues are equally weighted by the ESD so the impact of K signals vanishes when $T \to +\infty$. Now there are K signal eigenvalues and $T-K$ noise eigenvalues, and their respective characterization are different.[4] Noise eigenvalues are still distributed according to a TW distribution. The situation is a little bit more complicated for the signal eigenvalues since the characterization depends on a threshold τ. In the case of a white Gaussian noise, when no whitening transfomation in applied, the threshold is defined as $\tau = 1+\sqrt{c}$ [16]. The threshold takes into account the c_1 ratio in Property 6.3, when dealing with a colored noise [22]. When the i^{th} largest signal eigenvalue $\hat{\lambda}_i$ is strictly higher than τ, the signal eigenvalue converges to a limit different from that in the signal-free case [16]. In that case, the signal eigenvalue is Gaussian distributed, i.e.,

$$\mathrm{P}\left[\frac{\hat{\lambda}_i - \mu_i}{\sigma_i} \leq x\right] \to G(x)$$

[4]This case is often referred to as a "spiked" model in RMT.

where

$$\mu_i = \lambda_i(1 + \frac{c}{\lambda_i - 1})$$
$$\sigma_i = \frac{\lambda_i}{\sqrt{T}}\sqrt{1 - \frac{c}{(\lambda_i - 1)^2}}$$

and

$$G(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)dy$$

In Fig. 5, the pdf of the smallest signal eigenvalue is represented. Otherwise, when the i^{th} largest signal eigenvalue $\hat{\lambda}_i$

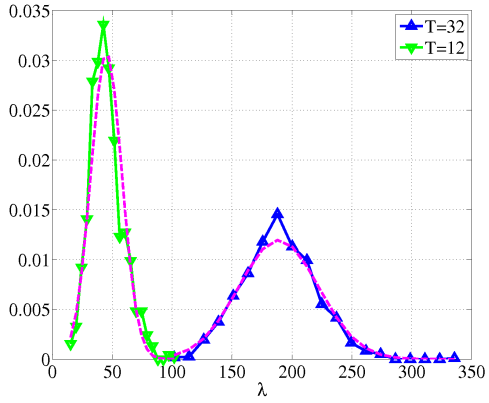


Fig. 5. Probability density function of the smallest signal eigenvalue of a matrix $\widehat{\mathbf{R}}^{\dagger}$ in the signal-bearing case with $K = 4$, i.e. $\hat{\lambda}_5$, with $m = 61440$, a noise variance of $\sigma^2 = 0.1$, and different values for T. The solid curves with markers are the simulation results and the dotted curves are the Gaussian distribution.

is strictly lower that τ, the signal eigenvalue is distributed according to a TW distribution, i.e.,

$$\mathrm{P}\left\{\frac{\log(\frac{m}{N}\hat{\lambda}_i) - \mu(T-i, m, N)}{\sigma(T-i, m, N)} \leq x\right\} \to TW_{\mathbb{C}}(x)$$

Note that there is a specific phenomenon when the i^{th} largest signal eigenvalue exactly equals to the threshold [24]. Hence, when $K \ll T$, the signal eigenvalues exhibit, on rescaling, fluctuations described by the TW distributions when there are strictly below the threshold τ. So the distributions obtained for the signal-free case ($K = 0$) closely approximate the distribution of the signal eigenvalues. These results have an impact on the design of CM estimation techniques since signal eigenvalues strictly below τ are considered as being noise eigenvalues. Note that adding observations, i.e., increasing T, cannot be an option to tackle this problem since τ grows with T. However, when the signal eigenvalues are above the threshold, a reliable estimation of the CM is possible.

IV. Collision Multiplicity Estimation Techniques

We first present the CM estimation method based on the distribution of noise eigenvalues. Then we review two approaches based on the distribution of signal eigenvalues.

A. A method based on the distribution of noise eigenvalues

We assume that T observations are available at the destination node. The SCM $\widehat{\mathbf{R}}^{\dagger}$ is computed from (4). Then, an eigenvalue decomposition of $\widehat{\mathbf{R}}^{\dagger}$ is performed and the resulting eigenvalues are sorted in descending order $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_T$. The method relies on the property that the i^{th} largest noise eigenvalue is TW distributed, provided that all signal eigenvalues are located above the detectability threshold. So for every $l = 0, 1, \ldots, T-1$, we test the null hypothesis that "there are exactly l colliding signals" against the alternative hypothesis that there are "at least $l+1$ colliding signals". The test is performed by computing the test statistic $S_\lambda(l)$

$$S_\lambda(l) = \frac{\log(\frac{m}{N}\hat{\lambda}_{l+1}) - \mu(T-l, m, N)}{\sigma(T-l, m, N)} \tag{8}$$

and comparing it to a threshold τ_α . The threshold τ_α is defined as

$$\tau_\alpha = TW_{\mathbb{C}}^{-1}(1-\alpha) \tag{9}$$

where α is some significance level. More precisely, a significance level α is set so that the probability that the null hypothesis is detected by chance is α. So the null hypothesis is accepted when we have

$$S_\lambda(l) < \tau_\alpha \tag{10}$$

The sequence of tests begins with $l = 0$. If the test $S_\lambda(0)$ is passed, i.e., when $S_\lambda(0) < \tau_\alpha$, there are no colliding users. Otherwise, there is at least one signal. The procedure proceeds to the next step with $l = 1$. The tests are performed subsequently until the test is passed. When the test is passed at step $l = q$, the eigenvalue $\hat{\lambda}_{q+1}$ is detected as being TW distributed. So $\hat{\lambda}_{q+1}$ is a noise eigenvalue, and hence there are exactly q colliding signals. Once the number of colliding users has been determined, the destination node stops the CM estimation process and proceeds to the next processing block. The procedure is summarized in Algorithm 1. The

Algorithm 1 TWIT algorithm

```
Compute R̂†
Perform the eigenvalue decomposition of R̂†
Sort the eigenvalues λ̂_i, i = 1, ..., T of R̂†
K̂_TWIT ← 0 and Test ← False
while Test = False and K̂_TWIT < T do
   μ ← μ(T − K̂_TWIT, m, N)
   σ ← σ(T − K̂_TWIT, m, N)
   Test ← {σ^{-1}[log(mλ̂_{K̂_TWIT+1}/N) − μ] < τ_α}
   if Test = False then
      K̂_TWIT ← K̂_TWIT + 1
   else
      break
   end if
end while
```

mean $\mu(x, y, z)$ and the standard deviation $\sigma(x, y, z)$ in the algorithm are defined in Property 6.4. Note that this criterion has been originally designed for arbitrary (or colored) noise [22]. That is the reason why the algorithm uses the eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$.

B. Methods based on the distribution of signal eigenvalues

We present two CMETs that are all based on the Gaussian distribution of signal eigenvalues. We first present the well-known MDL criterion, and then address another criterion based on recent advances in RMT.

1) The MDL criterion: The MDL criterion has been defined in [17]. This criterion has been used for decades in the area of signal array processing, and other related domains. The MDL estimator $\hat{K}_{\text{MDL}}$ for the CM K is defined as

$$\hat{K}_{\text{MDL}} = \underset{k=1,\ldots,T}{\operatorname{argmin}} \{\text{MDL}(k)\}$$

where

$$\text{MDL}(k) = -m(T-k)\log\left[\frac{g(k)}{a(k)}\right] + \frac{1}{2}k(2T-k)\log(m)$$

where

$$g(k) = \prod_{i=k+1}^{T} \hat{\lambda}_i^{\frac{1}{T-k}} \qquad a(k) = \frac{1}{T-k}\sum_{i=k+1}^{T}\hat{\lambda}_i$$

where the $\hat{\lambda}_i$ denote the eigenvalues of $\widehat{\mathbf{R}}^{\dagger}$ with $1 \leq i \leq T$, ordered in descending order. This estimator is consistent in the $m \to \infty$ sense. One of the reason why the MDL criterion has been widely used over the past two decades comes from its robustness to model mismatch, in particular when the underlying assumptions of snapshots and noise Gaussianity can be relaxed [25]–[27].

2) The criterion based on recent results on signal eigenvalues: This new criterion has been designed using the last results in RMT [16]. We denote this estimator SEMOS (Signal Eigenvalue-based Model Order Selection). The estimator $\hat{K}_{\text{SEMOS}}$ for the CM K is defined as, for complex data

$$\hat{K}_{\text{SEMOS}} = \underset{k=0,\ldots,T-1}{\operatorname{argmin}} \left\{ \left[\frac{1}{2}(\frac{1}{c})^2 q_k^2\right] + 2(k+1) \right\}$$

The test statistic q_k is defined as

$$q_k = \left[(T-k)\frac{g_s(k)}{a_s(k)} - (1+c)\right] \times T \tag{11}$$

where

$$g_s(k) = \sum_{i=k+1}^{T}\hat{\lambda}_i^2 \qquad a_s(k) = (\sum_{i=k+1}^{T}\hat{\lambda}_i)^2$$

V. Simulation Results

We study the performance of three estimation techniques: the MDL criterion, the TWIT, and the SEMOS estimator. The methods are evaluated in the context of Rayleigh block-fading channels. The channel coefficients A_{ij} in (1) are circularly symmetric Gaussian coefficients with zero mean and power unity and the coefficients are randomly changing from one observations to another. User stations are transmitting OFDM signals. The signals are constructed according to the

TABLE I
OFDM SIGNAL PARAMETERS

Modulation	BPSK
Number of sub-carriers N_{sub}	1024
Guard Interval GI	1/4
Number of OFDM symbols per OFDM block N_{OFDM}	48

IEEE 802.11 standard [28] and use the signal parameters listed in Table I.

From the signal parameters, we get that the number of samples per OFDM frame is $m = (1 + \text{GI}) \times N_{\text{sub}} \times N_{\text{OFDM}} = 61440$. For the sake of simplicity, we set that $N = m$, i.e., $N = 61440$. The performance of estimation techniques have been evaluated over 10,000 Monte Carlo trials ($N_{\text{Simu}} = 10000$) using the MATLAB software.

The three estimation techniques are compared in Fig. 6 to 10 in the case of a white Gaussian noise. The results have been obtained for both a variable T and a variable SNR. The significance level α for the TWIT is set to 0.01 [22]. Figures 6 and 7 represent the simulation results for $K = 3$ and $K = 4$ respectively. The simulation results are similar with $K = 3$ and $K = 4$ so we focus on the latter case in the following. When the number of observations and the SNR are increasing, the curves are converging toward the same position so it is difficult to compare them. So we choose instead to compare the estimation techniques with respect to the MSE ϵ^2 between the estimates $\hat{K}$ and K

$$\epsilon^2 = \frac{1}{N_{\text{Simu}}} \sum_{i=1}^{N_{\text{Simu}}} |\hat{K}_i - K|^2 \qquad (12)$$

where $\hat{K}_i$ is the estimate of K for the i^{th} trial. This is shown in Fig. 8 to 10. Since the curves are plotted on a log scale, only non-zero values are plotted. So when the experimental MSE is zero, the corresponding point of the curve is not represented.

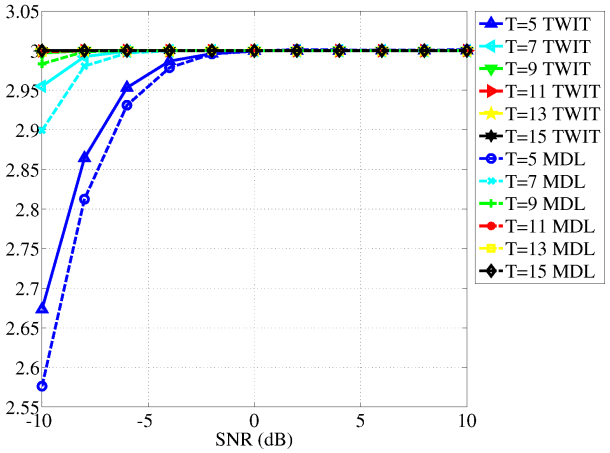


Fig. 6. Estimates of $K = 3$ with the TWIT and the MDL criterion for a variable number of snapshots T, $5 \leq T \leq 15$, and different SNR values.

The simulation results show that the estimation techniques perform better at high SNR and when T increases. The number of observations should be higher than almost three times the CM K in order to get a relative MSE lower than 10%. Similarly, the SNR should be higher than 0 dB in order to get a similar performance level, for any values of T. The two RMT-based techniques, i.e., the SEMOS estimator and the TWIT, outperforms the MDL criterion. However, the differences between the simulation curves decrease when the SNR or T is increasing. Moreover, the SEMOS estimator outperforms the TWIT. This can be due to the fact that the former technique used all the eigenvalues in the computation of the estimator while the latter only uses one eigenvalue to detect the number of signals. Note, however, that the TWIT has not been designed with the purpose of outperforming the SEMOS technique. Rather, the TWIT has been designed to cope with colored Gaussian noises. This is illustrated in the next figure.

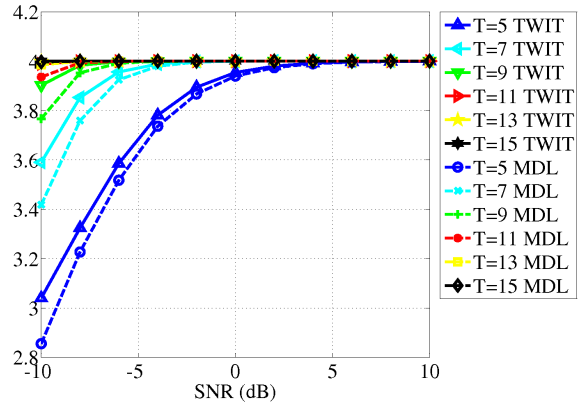


Fig. 7. Estimates of $K = 4$ with the TWIT and the MDL criterion for a variable number of snapshots T, $5 \leq T \leq 15$, and different SNR values.

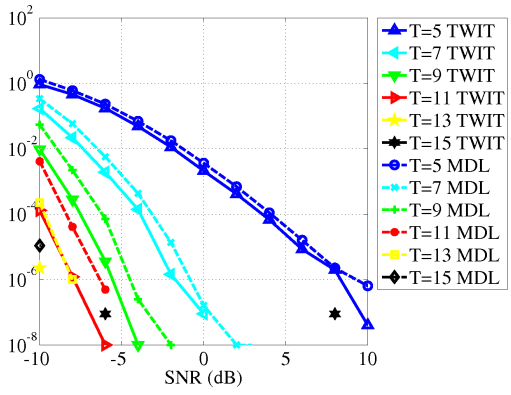


Fig. 8. MSE of two estimation techniques (TWIT and the MDL criterion) for $K = 4$, a variable number of snapshots T, $5 \leq T \leq 15$, and different SNR values.

Fig. 11 represents the performance of both the SEMOS

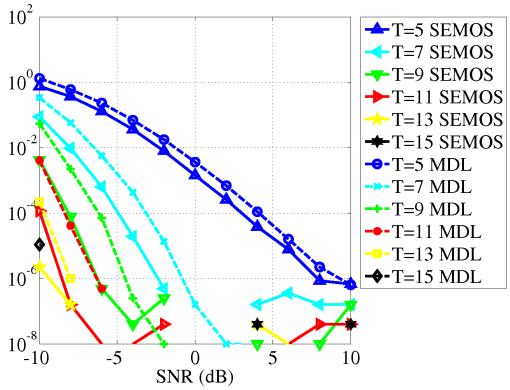


Fig. 9. MSE of two estimation techniques (the SEMOS estimator and the MDL criterion) for $K = 4$, a variable number of snapshots T, $5 \leq T \leq 15$, and different SNR values.

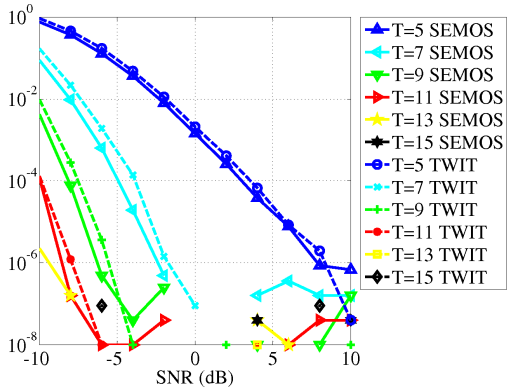


Fig. 10. MSE of two estimation techniques (the SEMOS estimator and the TWIT) for $K = 4$, a variable number of snapshots T, $5 \leq T \leq 15$, and different SNR values.

estimator and the TWIT. The simulation parameters are the same as the ones that have been used in previous simulations. The noise is now a colored noise. For this purpose, a white Gaussian noise with variance unity is passed through a filter with N_f coefficients f_i, $1 \leq i \leq N_f$ given in Table II.

We have that

$$\sum_{i}^{N_f} |f_i|^2 = 1$$

so the colored Gaussian noise at the output of the filter is also unit variance. The simulation results in Fig. 11 show that the SEMOS estimator is unable to estimate correctly the CM while the TWIT is still performing well.

TABLE II
FILTER COEFFICIENTS

0.227	0.460	0.688	0.460	0.227

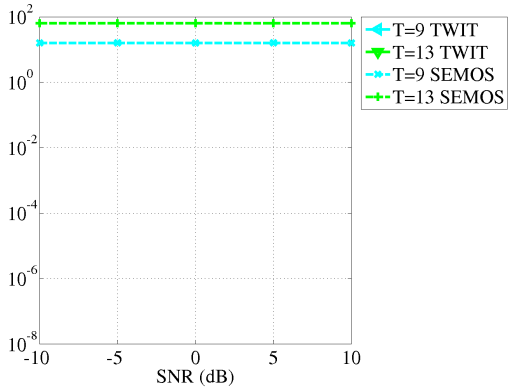


Fig. 11. MSE of two estimation techniques (the SEMOS estimator and the TWIT criterion) for $K = 4$, a variable number of snapshots T, $5 \leq T \leq 15$, and different SNR values, in the presence of a colored noise.

A. Discussion

The simulation results show that the estimation techniques perform better when the number of observations T is much larger than the number of signals K. This is in stark contrast with previous studies, such as the ones implemented in [7] and [8] where T is set to a value not greater than $K + 2$. In other approaches [11], [12], a single observation is required to perform the MUD of the colliding users. These references do not explicitly claim that the number of signals is estimated using only one observation. In this context, the blind separation technique designed in [13], is implemented and the number of sources (users) is assumed to have been estimated using an MDL-like criterion.

VI. CONCLUSION

In this paper, two new CM estimation techniques, the TWIT and the SEMOS estimator, have been proposed. The methods are based on eigenvalue statistics. In the TWIT, eigenvalues are tested in descending order, from the largest to the lowest. The first eigenvalue $\hat{\lambda}_q$ that is considered as being Tracy-Widom distributed allows the CM estimation by $\hat{K} = q - 1$. In the SEMOS estimator, a criterion is computed just as in the MDL criterion approach. These CM estimation techniques have been shown to outperform the typical MDL criterion. The SEMOS estimator outperforms the TWIT in the presence of a white Gaussian noise but it is inefficient in the presence of a colored noise. Moreover, simulation results have shown that a large number of snapshots T is needed in order to allow a good estimation of K in terms of MSE. Furthermore, the number of snapshots must be significantly higher than the number of colliding users K ($T \gg K$). These settings are similar to the settings that are used in MOS techniques for signal array processing.

The impact of these results is twofold. First, some CR techniques such as the network-assisted diversity multiple access (NDMA) [7], [8] cannot be implemented in IEEE 802.11

networks notably because these CR techniques are based on the assumption that T can be made as small as $K+1$ or $K+2$. Second, some multiple packet reception (MPR) protocols for IEEE 802.11 networks that use the blind user separation in [13] appear to be rather questionable since they assume that the CM estimation can rely on a single observation of collided request-to-send (RTS) frames. Even if the destination node is equipped with four antennas ($T = 4$), our simulation results have shown that the receiver at the destination node needs many more snapshots in order to provide a good estimation of K. This paper has pointed out a strong constraint in the design of MPR techniques. It revealed that a single observation of the colliding signals is far from providing enough information to estimated the number of colliding nodes.

Further investigations are now needed in order to fully characterized the performance of the proposed CM estimation techniques in typical operating conditions. The obtained results will allow the implementation of these techniques in current or future standards.

APPENDIX

A. Complex Wishart Matrices

Property 6.1: Consider a $T \times m$ matrix $\mathbf{X}$ with m samples drawn from a T-dimensional complex Gaussian law with zero mean and covariance matrix $\mathbf{R_X}$, denoted $\mathbb{C}\mathcal{N}_T(\mathbf{0}, \mathbf{R_X})$. The random matrix $\mathbf{X}\mathbf{X}^{\mathrm{H}}$ is a T-variate complex Wishart matrix with m degrees of freedom [20]. This is denoted

$$\mathbf{X}\mathbf{X}^{\mathrm{H}} \sim \mathbb{C}W_T(m, \mathbf{R_X})$$

where $\mathbb{C}W_T(m, \mathbf{R_X})$ denotes complex Wishart law, parameterized accordingly.

B. Empirical Spectral Distribution (ESD) of a Matrix

The ESD of an $T \times T$ Hermitian matrix $\mathbf{X}$, denoted $F^{\mathbf{X}}(x)$, is the distribution function of the eigenvalues of $\mathbf{X}$, i.e., for $x \in \mathbb{R}$

$$F^{\mathbf{X}}(x) = \frac{1}{T}\sum_{i=1}^{T} 1_{\{x, \lambda_i \leq x\}}(x) \tag{13}$$

where $\lambda_1, \ldots, \lambda_T$ are the population eigenvalues of $\mathbf{X}$ and $1_A(x)$ is the indicator function of the set A, i.e., $1_A(x) = 1$ if $x \in A$, and $1_A(x) = 0$ otherwise. The Hermitian property is required to ensure that all eigenvalues of $\mathbf{X}$ belong to the real line.

Property 6.2: Let $\mathbf{X} \in \mathbb{C}^{T \times m}$ be a random matrix with independent and identically distributed (iid) entries X_{ij} such that X_{ij} has zero mean and variance λ. As $T, m \to \infty$ with $T/m \to c \in (0, \infty)$, the ESD of

$$\mathbf{B} = \frac{1}{m}\mathbf{X}\mathbf{X}^{\mathrm{H}}$$

converges almost surely to a non-random distribution function with density $f_c(x)$ given by

$$\begin{aligned} f_c(x) &= (1 - \frac{1}{c})^+ \times \delta(x) \\ &+ \frac{1}{2\pi\lambda x c}\sqrt{(x-a)^+(b-x)^+} \end{aligned} \tag{14}$$

with $a = (1 - \sqrt{c})^2$, $b = (1 + \sqrt{c})^2$, $\delta(x) = 1_{\{0\}}(x)$ and, for $x \in \mathbb{R}$, $x^+ = \max(0, x)$.

The probability density function (pdf) $f_c(x)$ is the Marčenko-Pastur density.

Property 6.3: Let $\mathbf{B} \in \mathbb{C}^{T \times m}$ be a random matrix defined as

$$\mathbf{B} = \frac{1}{m}\mathbf{T}^{1/2}\mathbf{X}\mathbf{X}^{\mathrm{H}}\mathbf{T}^{1/2} \tag{15}$$

where $\mathbf{X} \in \mathbb{C}^{T \times m}$ is a random matrix with iid entries that have zero mean and power unity, and where $\mathbf{T}$ is an $T \times T$ Hermitian non negative definite matrix. As $T, m \to \infty$ with $T/m \to c \in (0, \infty)$, the ESD of $\mathbf{B}$ converges almost surely to a non-random distribution function with density $\widetilde{f}_c(x)$ given by

$$\begin{aligned} \widetilde{f}_c(x) &= (1 - \frac{1}{c})^+ \times \delta(x) \\ &+ \frac{1 - c_1}{2\pi x(xc_1 + c)}\sqrt{(x-b_1)^+(b_2-x)^+} \end{aligned} \tag{16}$$

with

$$\begin{aligned} b_1 &= \left[\frac{1 - \sqrt{1 - (1-c)(1-c_1)}}{1 - c_1}\right]^2 \\ b_2 &= \left[\frac{1 + \sqrt{1 - (1-c)(1-c_1)}}{1 - c_1}\right]^2 \end{aligned}$$

where $T/N \to c_1 \in (0, 1)$ as $T \to +\infty$.

The probability density function (pdf) $\widetilde{f}_c(x)$ is a modified Marčenko-Pastur density.

C. Distribution of the Largest Eigenvalue of Wishart Matrices

We now investigate the distribution of the largest eigenvalue of Wishart matrices. Consider $\mathbf{A} \sim \mathbb{C}W_T(m, \mathbf{R_b})$ and $\mathbf{B} \sim \mathbb{C}W_T(N, \mathbf{R_b})$, two independent random matrices where $m, N > T$. We consider the generalized eigenproblem constructed from $\mathbf{A}$ and $\mathbf{B}$

$$\mathbf{A}v = \theta(\mathbf{A} + \mathbf{B})v \tag{17}$$

where v denotes the eigenvector corresponding to the eigenvalue θ. This is referred to as a double Wishart setting [20].

Property 6.4: The largest eigenvalue satisfying (17), denoted $\lambda_1^{(D)}$, is distributed according to a Tracy-Widom (TW) distribution when $m, N \to \infty$ as $T \to \infty$ with $m, N > T$, i.e.,

$$\mathrm{P}\left[\frac{W(\lambda_1^{(D)}) - \mu(T, N, m)}{\sigma(T, N, m)} \leq x\right] \to TW_{\mathbb{C}}(x) \tag{18}$$

where $W(\theta)$ denotes the logit transformation of θ, i.e., $W(\theta) = \log[\theta/(1-\theta)]$ and $TW_{\mathbb{C}}(x)$ is the TW distribution function for complex data.

The centering $\mu(T,N,m)$ and rescaling $\sigma(T,N,m)$ in (18) are parameterized as follows

$$\begin{aligned}
\beta &= \min(N,T) \\
\gamma &= m-T \\
\delta &= |N-T| \\
\mu(T,m,N) &= (\frac{u_\beta}{\tau_\beta}+\frac{u_{\beta-1}}{\tau_{\beta-1}})(\frac{1}{\tau_\beta}+\frac{1}{\tau_{\beta-1}})^{-1} \\
\sigma(T,m,N) &= 2(\frac{1}{\tau_\beta}+\frac{1}{\tau_{\beta-1}})^{-1} \\
\sin^2(\gamma_\beta/2) &= (\beta+1/2) \\
&\times (2\beta+\gamma+\delta+1)^{-1} \\
\sin^2(\phi_\beta/2) &= (\beta+\delta+1/2) \\
&\times (2\beta+\gamma+\delta+1)^{-1} \\
\tau_\beta^3 &= 16(2\beta+\gamma+\delta+1)^{-2} \\
&\times \sin^{-2}(\phi_\beta+\gamma_\beta) \\
&\times \sin^{-1}(\phi_\beta)\sin^{-1}(\gamma_\beta) \\
u_\beta &= 2\log[\tan(\frac{\phi_\beta+\gamma_\beta}{2})]
\end{aligned}$$

Note that the covariance matrix $\mathbf{R_b}$ has no effect on the distribution of the eigenvalue $\lambda_1^{(D)}$. A proof of Property 6.4 can be found in [21] for $\mathbf{R_b}=\mathbf{I}_T$. We now rewrite (17) as

$$\mathbf{A}\mathbf{B}^{-1}v = \frac{\theta}{1-\theta}v \tag{19}$$

and derive a new property for the following generalized eigenproblem

$$\mathbf{A}v = \theta\mathbf{B}v \tag{20}$$

This is referred to as a single Wishart setting. From (19) and (18), we get the following property.

Property 6.5: The largest eigenvalue satisfying (20), denoted $\lambda_1^{(S)}$, is distributed according to a TW distribution when $m,N\to\infty$ as $T\to\infty$ with $m,N>T$, i.e.,

$$\mathrm{P}\left\{\frac{\log[\lambda_1^{(S)}]-\mu(T,N,m)}{\sigma(T,N,m)}\le x\right\}\to TW_{\mathbb{C}}(x)$$

REFERENCES

[1] B. Escrig, "Estimation of Collision Multiplicities in IEEE 802.11-based WLANs," in *Proc. IARIA Eleventh International Conference on Networks (ICN)*, 2012.

[2] P. V. Lang Tong, V. Naware, "Signal processing in random access," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 29–39, 2004.

[3] M. Z. A. A. F. Borgonovo, L. Fratta, "Capture division packet access: a new cellular access architecture for future PCNs," *IEEE Communications Magazine*, vol. 34, no. 9, pp. 154–162, 1996.

[4] D. L. S. Y. Park, D. Park, "On Scheduling for Multiple-Antenna Wireless Networks Using Contention-Based Feedback," *IEEE Transactions on Communications*, vol. 55, no. 6, pp. 1174–1190, 2007.

[5] P. T. M. Durvy, O. Dousse, "Self-Organization Properties of CSMA/CA Systems and Their Consequences on Fairness," *IEEE Transactions on Information Theory*, vol. 55, no. 3, pp. 931–943, 2009.

[6] S. Verdú, "Multiuser Detection." Cambridge Press, 1998.

[7] R. Zhang, N. D. Sidiropoulos, and M. K. Tsatsanis, "Collision Resolution in Packet Radio Networks Using Rotational Invariance Techniques," *IEEE Transactions on Communications*, vol. 50, no. 1, pp. 146–155, 2002.

[8] B. Özgül and H. Deliç, "Wireless Access with Blind Collision-Multiplicity Detection and Retransmission Diversity for Quasi-Static Channels," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 858–867, 2006.

[9] A. P. R. Lin, "A new wireless network medium access protocol based on cooperation," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, 2005.

[10] R. Z. G. Dimic, N.D. Sidiropoulos, "Medium access control - physical cross-layer design," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 40–50, 2004.

[11] P. X. Zheng, Y. J. Zhang, and S. C. Liew, "Multipacket Reception in Wireless Local Area Networks," in *Proc. IEEE International Conference on Communications (ICC)*, 2006.

[12] W. L. Huang, K. B. Letaief, and Y. J. Zhang, "Cross-Layer Multi-Packet Reception Based Medium Access Control and Resource Allocation for Space-Time Coded MIMO/OFDM," *IEEE Transactions on Wireless Communications*, vol. 7, no. 9, pp. 3372–3384, 2008.

[13] S. Talwar, M. Viberg, and A. Paulraj, "Blind Separation of Synchronous Co-Channel Digital Signals Using an Antenna Array. Part I. Algorithms," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1184–1197, 1996.

[14] P. Stoica and Y. Selèn, "Model-Order Selection : A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[15] Q. Ding and S. Kay, "Inconsistency of the MDL: On the Performance of Model Order Selection Criteria with Increasing Signal-to-Noise Ratio," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 1959–1969, 2011.

[16] R. R. Nadakuditi and A. Edelman, "Sample Eigenvalue Based Detection of High-Dimensional Signals in White Noise Using Relatively Few Samples," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2625–2638, 2008.

[17] M. Wax and T. Kailath, "Detection of Signals by Information Theoretic Criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[18] M. S. Bartlett, "A note on the multiplying factors for various χ^2 approximations," *J. R. Stat. Soc.*, vol. 16, p. 296298, 1954.

[19] D. N. Lawley, "Tests of significance of the latent roots of the covariance and correlation matrices," *Biometrica*, vol. 43, p. 128136, 1956.

[20] I. M. Johnstone, "High Dimensional Statistical Inference and Random Matrices," in *Proceeding of the International Congress of Mathematicians*, 2006.

[21] ——, "Multivariate analysis and Jacobi ensembles: Largest eigenvalue, TracyWidom limits and rates of convergence," vol. 36, no. 6, pp. 2638–2716, 2008.

[22] R. R. Nadakuditi and J. W. Silverstein, "Fundamental Limit of Sample Generalized Eigenvalue Based Detection of Signals in Noise Using Relatively Few Signal-Bearing and Noise-Only Samples," *IEEE Transactions on Signal Processing*, vol. 4, no. 3, pp. 468–480, 2010.

[23] P. O. Perry and P. J. Wolfe, "Minimax Rank Estimation for Subspace Tracking," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 504–513, 2010.

[24] J. Baik, G. B. Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability*, vol. 33, no. 5, pp. 1643–1697, 2005.

[25] G. Xu, R. H. Roy, and T. Kailath, "Detection of number of sources via exploitation of centro-symmetry property," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 102–112, january 1994.

[26] P. Stoica and M. Cedervall, "Detection tests for array procesing in unknown correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 45, no. 9, september 1997.

[27] E. Fishler and H. V. Poor, "Estimation of the number of sources in unbalanced arrays via information theoretic criteria," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3543–3553, september 2005.

[28] "IEEE 802.11n standard, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Amendment 5: Enhancements for Higher Throughput," IEEE Computer Society, Tech. Rep., 2009.

Evaluation of Opportunistic Routing Algorithms on Opportunistic Mobile Sensor Networks with Infrastructure Assistance

Viet-Duc Le, Hans Scholten, and Paul Havinga
Pervasive Systems (PS), Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
{*v.d.le, hans.scholten, p.j.m.havinga*}*@utwente.nl*

***Abstract*—Recently the increasing number of sensors integrated in smartphones, especially the iPhone and Android phones, has motivated the development of routing algorithms for Opportunistic Mobile Sensor Networks (OppMSNs). Although there are many existing opportunistic routing algorithms, researchers still have an ambiguous understanding of how these schemes perform on OppMSNs with heterogeneous architecture, which comprises various kinds of devices. In this work, we investigate the performance of well-known routing algorithms in realistic scenarios. To this end, we propose a heterogeneous architecture including fixed infrastructure, mobile infrastructure, and mobile phones. The proposed architecture focuses on how to utilize the available, low cost short-range radios of mobile phones for data gathering and dissemination. We also propose new realistic mobility models and metrics. Selected routing protocols are simulated and evaluated with the proposed heterogeneous architecture, mobility models, and transmission interfaces under various constraints, such as limited buffer size and time-to-live (TTL). Results show that some protocols suffer long TTL, while others suffer short TTL. We further study the benefit of fixed infrastructure in network performance, and learn that most of the opportunistic routing algorithms cannot benefit from the advantage of fixed infrastructure since they are designed for mobile nodes. Finally, we show that heterogeneous architecture need heterogeneous routing algorithms, such as a combination of Epidemic, Spray and Wait, and context-based algorithms.**

Keywords-opportunistic sensor network; opportunistic routing; heterogeneous architecture; mobility model; smartphone.

I. Introduction

Mobile phones play an important role in sensor network applications during last few years. Measurements can be gathered with either user participatory, opportunity, or both. No matter by which, data gathering is particularly meaningful when performed by many phones simultaneously. To enhance data reliability and dissemination performances, heterogeneous architecture that consists of various kinds of sensor devices is necessary. In fact, this paper is an extension of Le et al. [1] to continue with evaluation of existing opportunistic routing algorithms on heterogeneous architecture .

Sensor networks, a large collection of nodes to collect the world's physical nature, have gone through various evolution phases as depicted in Figure 1. In Wired Sensor Networks, the deployment of sufficient sensors is often bounded by the cost of wiring. Later, Wireless Sensor Networks (WSNs) have been taken into consideration to replace the existing Wired Sensor Networks, since WSNs provide a wide range of context-awareness for real-time applications at low cost. A variety of sensor types with dense deployment forms a connected wireless mesh network via low power, short-range radios, collaborating to acquire and transmit the target data to sink nodes [2]. However, limited to unchangeable topology, WSNs cannot be applied for a variety of applications with different types of architecture or inaccessible areas. Therefore, Mobile Sensor Networks (MSNs) have been presented to facilitate the data collection of sensors. But still, the cost of deploying all kinds of such required sensors is considerably high in terms of time and money.

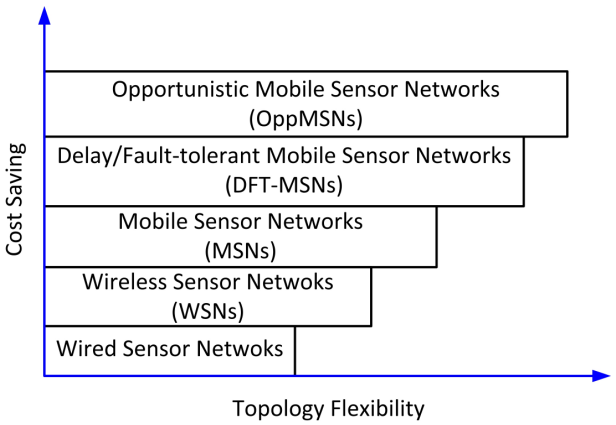


Figure 1. Evolution of sensor networks.

The next step in sensor networks is to enhance, or even replace, mobile nodes in MSNs with mobile phones. Thanks to developments in sensor technology, smart phones, such as the iPhone and Android-based phones, are equipped with a large number of sensors, including GPS, accelerometers, gyroscopes, proximity sensors and cameras. Even regular phones also have sensors: microphones, light sensors, and onboard radios. Not all mobile phones can access 3G mobile internet, especially when a disaster happens, for example, an earthquake or tsunami. But still mobile phones have the means to participate in the sensor network. This revolution

is termed as Delay/Fault-tolerant Mobile Sensor Networks (DTF-MSNs) [3]. Nevertheless, the architecture of DTF-MSNs lacks infrastructure, which most real-world applications often have. To this end, we propose new type of sensor networks, of which architecture consists of fixed infrastructure, mobile infrastructures, and mobile nodes (mainly smart phones). We term the new sensor networks as Opportunistic Mobile Sensor Networks (OppMSNs). Unlike Ad-hoc Networks (VANETs) use only RSUs, OppMSNs utilize a wide range of available devices to measure and disseminate data for specific tasks. For example, through WiFi or Bluetooth radio, mobile nodes can collaborate with nearby ones, cars, buses, laptops, and the existing infrastructure-based sensor networks for data gathering.

In addition, as requiring a contemporaneous end-to-end connectivity, traditional routing algorithms such as Ad-hoc On-Demand Distance Vector (AODV) [4] or Dynamic Source Routing (DSR) [5], which can be used for MSNs, may perform poorly in scenarios where the communication paths are disrupted because of the sparse and mobility of sensor nodes. Opportunistic routing algorithms with the store-carry-forward paradigm, which can be applied for OppMSNs, have been proposed in a number of recent studies to evaluate the performance of routing algorithms on data gathering [6]–[19]. However, these algorithms use either basic scenarios or simple simulation architectures that are still quite far from real-world applications.

This paper investigates the performance of existing opportunistic routing algorithms for OppMSNs by proposing heterogeneous architecture, mobility models and metrics. The architecture includes most of real-world sensor nodes such as Road Side Units (RSUs), buses, cars and pedestrians with unpredictable movement. To achieve a realistic setting, the architecture is mapped on a real city, the city of Enschede, The Netherlands. Buffer size and time-to-live of messages are limited. We also consider heterogeneous means of communication, especially WiFi and Bluetooth. In addition, two new models, together with available ones in The ONE simulator [20], will be implemented to make the investigation more realistic. By means of simulations, the proposed architecture and models are used for the comparison of a set of opportunistic routing protocols.

The paper has the following structure: related work is discussed in Section II. Section III presents the architecture, new mobility models and evaluation metrics. The simulations and an analysis of simulation results are the subject of Section IV, which covers evaluations of movement model, algorithms' performance, and RSUs' assistance. Based on the results, Section V gives possible directions for current and future research.

II. Related Work

In this paper, we evaluate performance in term of message delivery of most well-known opportunistic routing algorithms with OppWSNs that are essentially composed of the existing wireless ad-hoc sensor networks (RSUs) and the mobile sensor networks (flocks of mobile phones). The network can be characterized as intermittent connection and sparse mobility. Conventional routing protocols [4], [5] require contemporaneous end-to-end connectivity for a data packet to be delivered. In other words, if the destination is not available on the connected path, the packet delivery will fail and no further effort is taken to secure future transmission of the data. Consequently, routing protocols must be adapted for these new types of networks. Numerous opportunistic routing algorithms have been proposed in the last few years with different mechanisms [3], [6]–[15], [17]–[19], which can be generally categorized based on either the type of network (*without infrastructure* and *with infrastructure*) or the pre-known information of the networks (*Stochastic* and *Context-based*) as defined in [21]. These categorizations slightly overlap as depicted in Figure 2. If networks are sparse and most nodes possess unpredictable movement, the stochastic protocols are more appropriate. In our opinion, an algorithm that can combine advantages of stochastic and context-based approaches is most suitable for our considered networks since the global knowledge of fixed and mobile infrastructures perhaps improves the routing performance of mobile nodes, which have unpredictable movement.

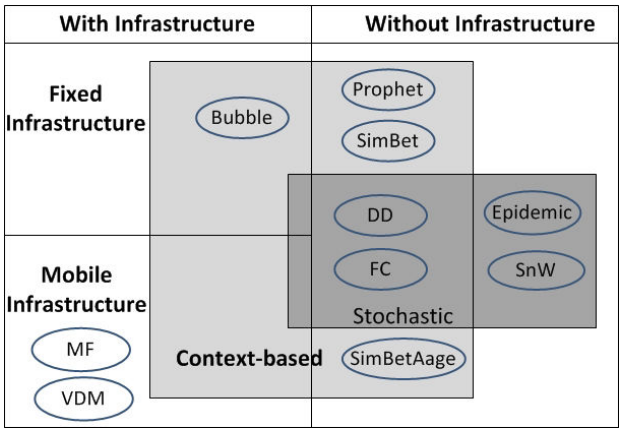


Figure 2. Categorizations of routing protocols in opportunistic networks.

A. Routing Without Infrastructure Assistance

Stochastic routing protocols deliver messages by simply disseminating them all over the network. Being passed from node to node, messages will be gradually delivered at the destination. Epidemic Routing [12] diffuses messages similar to the way virus/bacteria spread in biology. When encountering others, a node will replicate and broadcast the messages to them. These nodes that just received the messages will move to other places and continuously replicate and transmit messages to other nodes whenever they are in

range of communication. Though increasing the possibility of message delivery, the method results in flooding the network and rapidly exhausting available resources. Direct Delivery (DD) [13] only delivers the holding messages directly to the destination; therefore, DD saves huge amounts of resources but decreases significantly the delivery probability. Spray and Wait (SnW) [14] is a tradeoff between multi-copy scheme (Epidemic) and single-copy scheme (Direct Delivery) by finding an optimal number of message copies and dividing the message delivery process into two phases (*spray phase* and *wait phase*). First Contact (FC) [15] is a variant of single-copy scheme, which sends messages to the first encountered node or a random node if there are more than one.

In general, context-based protocols use information of historical contacts. Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET) [16] is a well-known Context-based routing protocol based on encounter. PRoPHET estimates the delivery predictability for each known destination at each node before passing a message. The estimation relies on the history of encounters between nodes. SimBet [17] uses historical contacts to calculate two metrics, similarity and betweenness. The similarity, which is calculated by how frequently a node and its destination have met, is meant of how socially connected such two nodes are. Meanwhile, the betweenness, which is calculated by how many nodes which a node has met, is meant of how interconnected a node is. However, if the utility metrics are equal, SimBet will prevent its forwarding behavior. To improve this flaw, BUBBLE [18] adds the knowledge of community structure to ensure message diffusion. Since the social knowledge varies over time, information used BUBBLE may be outdated. In addition, the betweenness may be useless if the message is near its destination. Motivated by this shortcoming, SimBetAge [19] is proposed.

B. Routing With Infrastructure Assistance

Data Mule [22] is designed to exchange messages between the close fixed infrastructure via mobile nodes with random movement. Conversely, Virtual Data Mule (VDM) [11], Message Ferrying (MF) and its variants [10], [23], [24] try to improve network performance by increasing the encounter probability via predefined movement. The ferries shuttle along the predefined routes in the dedicated region. Meanwhile, mobile nodes have tendency to move towards ferries to send messages. Such assumption makes the algorithms limited in specific scenarios with the majority are buses and bus travelers. In fact, these algorithms are entirely constrained by the route and time schedule of ferries. Without the route information, the algorithms will perform poorly.

C. Routing for OppMSNs

To our best knowledge, little attention has been given on how to apply above opportunistic routing algorithms on data dissemination in OppMSNs. DTF-MSN [3] shows a scheme to gather information in the Delay/Fault-Tolerant Mobile Sensor Network based on an improvement of Direct Delivery and Epidemic. The proposal consists of two key components: queue management and data transmission. Queue management decides the importance of messages, and data transmission decides the node with high delivery probability to send messages to. However, the scenario used to evaluate the proposal has only one mobility model, where both source and sink are mobiles nodes, and is far from realistic for the OppMSN application domain. Camara et al. [6] present a good mechanism for the distribution of messages, but the mechanism limits itself to vehicle-to-vehicle and infrastructure-to-vehicle. The work uses only the basic Epidemic routing and there is no comparison with other routing protocols. Recently, Keranen et al. [25] evaluate opportunistic networks with various mobility models and routing algorithms by using the ONE. Nevertheless, the used architecture does not include fixed infrastructure and the results only show the simulation speed.

Therefore, we are interested in investigating towards routing in OppMSNs, which consist of fixed, mobile infrastructures, and mobile nodes with unpredictable movement. Since algorithms are proposed for different optimization objectives under different constraints and scenarios, it is difficult to compare the performance of them all. In this paper, we only select the five most well-known and comparable to investigate. They are Epidemic [12], Direct Delivery (DD) [13], FirstContact (FC) [15], and PRoPHET [16], and Spray and Wait (SnW) [14].

We also improve the ONE simulator for simulations. The ONE includes several opportunistic routing algorithms and mobility models. Researchers can import their own maps and to configure the simulator with their own settings by many parameters, such as speed of mobility, message size, buffer size, and etc. Moreover, the ONE is an open source enabling researcher develop the tool for their own specific objectives.

III. Proposed Opportunistic Mobile Sensor Network (OppMSN)

Most traditional sensor applications are based on fixed and mobile wireless sensor networks, for which the availability of contemporaneous end-to-end connectivity is essential. However, the very recent innovation of mobile phones with different types of onboard sensors and available low power consumption radios has brought on a new interest of using mobile phone as the main part of sensor networks. The network becomes opportunistic, and mainly consists of the existing wireless ad-hoc sensor network and the mobile sensor network.

A. Architecture

The considered opportunistic network is separated into several regions based on communities as shown in Figure 3. In order to link these regions, each region has a base station equipped with long-range communication such as satellite, GSM, Internet. In addition, network architecture of a region consists of the following components: a fixed infrastructure, a mobile infrastructure (e.g. data mules), and mobile nodes.

- *Fixed infrastructure:* Road side units (RSUs) are deployed along main roads of the region. RSUs will be physically integrated in or fixed to the existing infrastructure, like lamppposts, GSM base station, or walls. RSUs form an ad-hoc wireless network, acting as a backbone, connecting mobile nodes with central servers or data sinks. The fixed infrastructure can also be used to disseminate information from central servers to the regions. The distance between RSUs is approximately 60 meters, using WiFi to build the network. There are two types of wireless interfaces for the RSUs, short-range Bluetooth and WiFi 802.11. Messages are transferred among RSUs through WiFi. The Wifi interface is also used to connect to buses, trams, cars, and smart phones. Bluetooth is designed for communication between RSUs and regular phones.
- *Mobile infrastructure:* Equipped with WiFi 802.11, buses and trams with known routes and known stops are considered as the mobile infrastructure in OppMSN applications. Since buses and trams move relatively fast, Bluetooth characterized by short-range (< 10 m) and low speed (< 2 Mbit/s) is not an appropriate option for buses and trams.
- *Mobile nodes:* The last component of the heterogeneous architecture consists of mobile phones (used by pedestrians) and a small portion of cars. There is no information of possible paths towards the sink because mobile phones, the majority of networks, move unpredictably. Mobile phones are classified into either smart phones or regular phones. Smart phones typically have both WiFi and Bluetooth interfaces, while regular phones have only Bluetooth. For the same reason buses and trams use WiFi only, cars are equipped with WiFi.

B. Architecture Performance Requirements

Depending on the physical characteristic, each of proposed components has a different degree of performance requirements such as reliability, throughput, latency, and electric power consumption. Fixed infrastructure has unlimited electric power supply, strong and stable signal strength, and large data storages. Therefore, latency and throughput are the most considerable performance requirements, and reliability and power consumption can be ignored. A message should be transferred as fast as possible via the ad-hoc connected network based on fixed infrastructure. Since

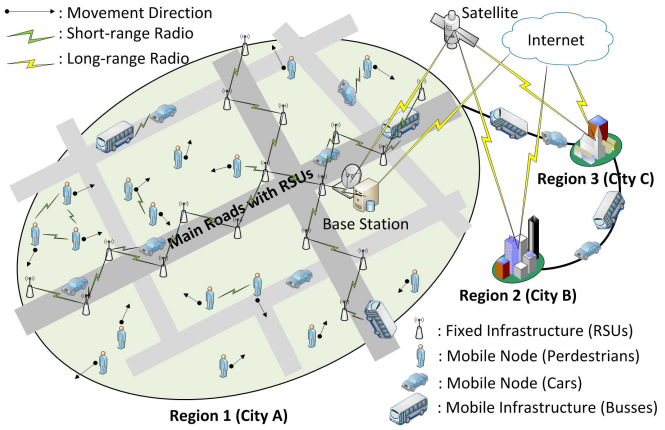


Figure 3. Architecture for Opportunistic Mobile Sensor Networks.

the RSU network is linear lines crossing each other at few points, the bottleneck phenomenon probably decreases throughput and increases latency.

Mobile infrastructure, such as buses and trams, has no constraint on power supply, signal strength, and storage capacity. Thus, mobile infrastructure also has no problem with reliability and power consumption. As buses or trams play a role as messengers shuttling between sources and sinks in the network, latency depends on velocity and distance significantly. In addition, mobile infrastructure may become a bottleneck point because many passengers try to connect to a bus or a tram. As a result, the throughput of mobile infrastructure needs to improve as well.

Since mobile phones suffer limited power supply and intermittent connectivity, power consumption and throughput are the most critical performance requirements. Reliability is another considerable performance requirement because mobile nodes are sparse and dynamic. That some people are not willing to turn on the wireless interfaces all the time also makes the network less reliable. Moreover, velocity and unpredictable movement patterns of mobile nodes deter obtaining low latency and high throughput.

C. Network Operations

When a node sends a message to the data sink (base station) by using an opportunistic protocol, the message is transferred towards the base station by the store-carry-forward paradigm. The message is stored in phones or vehicles, and then forwarded to other nodes during opportunistic contacts. The node receiving the message is either the base station, a car, a phone, or a RSU. The nodes, except the base station, continuously forward the message when in communication range of other nodes. Eventually, nodes may carry the message to the base station. If reaching a RSU, the message usually takes the paths based on connected RSUs to go to the base station.

RSUs with a large storage capacity also act as a relay node in the network. Messages are stored at RSUs for a period of time until they expire due to a limited time-to-live (TTL). In some cases, reliability of event detection is enhanced by aggregating data provided by other sensor devices. A mobile node perhaps receives messages from the fixed infrastructure and then forwards them to other nodes. As a result, a message containing event information will not only be transferred to the base station but also disseminated to nodes in network.

Buses and trams are not only message ferries as described in [24] but also gateways for passengers. Because the contact durations of mobile phones carried by passengers on a bus are quite long, messages may be fully exchanged among the passengers. Furthermore, these messages are stored at the bus and then disseminated to next passengers or delivered to the base station at the last bus stop.

When moving from one region to another, a mobile node will act as a gateway, transferring messages between regions. The transfer will be slow compared to using the fixed infrastructure. As the anticipated application domain is safety in public spaces, (emergency) messages should reach their destination as fast as possible.

D. Mobility Model

To increase the realism of the mobility model, we propose two additional models, Random Shortest Path Map Based Movement (RSPMBM) and Road Side Unit Placement. The new models, together with exiting Map Based Movement, Bus Traveler Movement and Bus Movement, are suitably applied for different types of sensor nodes. This approach represents the heterogeneous nature of reality, with Road Side Units, cars, buses and pedestrians.

We assume that a portion of mobile nodes represents pedestrians wandering around without any specific purpose. The existing Map Based Movement (MBM) provided by the ONE is likely the most suited. MBM is the Random waypoint movement with map-based constraints, in which a mobile node moves from one map node to another by selecting a neighboring map node randomly. This movement is repeated a randomly chosen number of times.

Naturally, people do not just wander around. They want to go somewhere for a purpose, using the shortest or fastest path possible. The choice between walking or taking the car is often decided by the Euclidean distance to the destination. These destinations are very diverse [26], ranging from points of interest in the public domain (e.g. restaurants, parks, offices) to the more private ones (e.g. friends, home, family). Therefore, we propose the new movement model, RSPMBM, to model the behavior of human-like mobility. A node selects an arbitrary destination within a predefined range and then moves along the shortest path. Euclidean distance ranges are configurable in a setting file, for example, the distance ranges can be set $[50, 500]$ and $[500, 5000]$ meters for pedestrians and cars, respectively. Remark that the minimum walking distance of a pedestrian is set to 50 m to ensure every node always travel a little.

It is reasonable to assume that a number of civilians, called as bus travelers, who prefer traveling by bus. Movements of bus travelers and buses are modeled by Bus Traveler Movement and Bus Movement that are also available in the ONE simulator, respectively. A bus traveler compares distance to the nearest bus stop with to the destination to decide whether to take a bus or not. A bus can carry many passenger and shuttles flowing its pre-defined route and timetable.

The new Road Side Unit Placement model is proposed for deploying RSUs on a map, along side roads with a certain distance between each other. The RSUs are stationary and form a wireless ad-hoc network or wireless sensor network.

E. Evaluation Metrics

Four metrics are used to evaluate the aforementioned performance requirements of different routing algorithms. Two of them are metrics implemented in the ONE: delivery probability and latency. Hop-count metric is no longer an informative metric to assess the delivery cost in time and distance in OppWSNs as it is used in connected ad-hoc WSNs. Instead, we define Delivery Speed and Delivery Cost for a more accurate evaluation.

- *Delivery Probability DP:* The total number of successfully delivered unique message, denoted by Q, divided by the total number of created unique messages, denoted by P. Each unique message is created at certain time, and has an unique identification number to be distinguished with others in the network.

$$DP = \frac{Q}{P}. \tag{1}$$

- *Latency (DL):* The average of delays between the moment that unique message i is originated, denoted by Ts_i, and the time when the first replicate of unique message i arrives at the destination, denoted by Td_i. The replicate is a copy of an unique message. The number of replicates depends on the methodology of the routing algorithm, single or multiple-copies.

$$DL = \frac{1}{Q}\sum_{i=1}^{Q}(Td_i - Ts_i). \tag{2}$$

- *Delivery speed (DS):* The average of speeds of the first replicate of unique message i that is sent from the origin to the destination. It is defined by the Euclidean distance, denoted by d_i, divided by latency.

$$DS = \frac{1}{Q}\sum_{i=1}^{Q}\frac{d_i}{Td_i - Ts_i}. \tag{3}$$

- *Delivery cost (DC):* The total number of unique messages including replicates, denoted by R, divided by the number of first replicates successfully arrived their destinations.

$$DC = \frac{P + R}{Q}. \quad (4)$$

Note that latency DL does not take the distance from origins to destinations into account like the delivery speed DS. Therefore, DL only a good metric in scenarios that origins of messages are uniform distributed.

To evaluate the proposed architecture and the proposed mobility model, we use the inter-contact time, first defined by Chaintreau et al. [27]. Inter-contact time is the time interval between two successive contacts of a pair of nodes, from the end of one contact to the next contact with the same node. Inter-contact time represents the frequency of opportunities for nodes to send packets to other nodes. The distribution of inter-contact time has an impact on the performances of different routing algorithms. [27] also shows that the inter-contact times are power-law distributed with the power-law exponent less than one.

IV. Simulation and Evaluation

In order to evaluate our proposed architecture and mobility model, a realistic simulation environment is set up, using a real city map. The results of running selected routing protocols are analyzed and compared to gain a better understanding on performances of existing routing protocols. From that, we may attain implications for future work. We use the ONE simulator [20] that is specially designed for opportunistic networks. It allows users to import maps, configure radios, message size, node speed, etc. The most advantage of the ONE is an open source so that we can flexibly develop new features for better simulation.

A. Environment Setup

The simulation uses the center of the city of Enschede as a realistic setting. In the center of the map, there is the central bus station. The map shown in Figure 4 takes up approximately 3500 by 3000 meters and is exported from Openstreetmap.org. To this map several layers, as submaps, are added for RSUs, roads for cars, paths for pedestrians and routes for buses. RSUs are positioned at the outer and inner ringroads, and four main roads radiating from the center. Cars are restricted to roads, but pedestrians may roam everywhere. Buses follow routes from the real city bus system. Roads in the ONE simulation have zero width. To overcome this limitation, roads are defined by two parallel routes as the lanes of a real road. In this way, communication with vehicles or pedestrians at both sides of the road is more realistic.

The simulation is carried out with 336 RSUs manually fixed on main roads, 50 cars, and 600 pedestrians moving around inside the city. The initial position of cars and pedestrians is randomly distributed. There are quite many bus lines in the city but only four are chosen because others have routes overlapping the RSU lines. Since RSUs can transfer messages to the base station much faster than buses do, buses that run along RSU lines have small contributions to the message delivery. Each bus line has two buses shuttling their routes. Since our basic goal is to investigate the contribution of pedestrians in disseminating data, only a small portion of cars, 50 over 650 mobile nodes, are simulated in the simulation. We also assume that the speed of pedestrians remains almost constant, $0.5 - 1.5$ m/s. Therefore, the mobility speed has a minor effect on performance results.

Since our proposed architecture also aims to reduce the use of mobile services for message exchange, we only consider available short-range interfaces, particularly Bluetooth and WiFi. All mobile phones have Bluetooth Version 2.0 at 2 Mbit/s net data rate with 10 m radio range, while smart phones have only WiFi interface at net data rate of 10 Mbit/s with 60 m radio range. We assume that fifty percent of pedestrians own smart phones and the rest uses normal phone. RSUs have both interfaces. The remaining nodes, cars and buses, use WiFi only, because they move at speeds that make Bluetooth communication unrealistic.

From the 600 pedestrians, 500 move with a purpose, while 100 are just strolling. Because cars likely possess predetermined routes, RSPMBM would be most suited. Buses follow fixed routes with predefined stops, and are modeled with the Bus Movement mobility model. Finally, pedestrians in buses are modeled with the Bus Traveler Movement model.

Data dissemination in the above heterogeneous scenario is simulated with a number of opportunistic routing protocols: Epidemic [12], Direct Delivery (DD) [13], FirstContact (FC) [15], and PRoPHET [16], and Spray and Wait (SnW) [14] with the number of copies (n) to be 6. This setting value is default in the ONE simulator. Since Message Ferry (MF) [24] is only useful for buses to transfer messages among base stations of cities, in our simulation with a single city, buses are just considered as a vehicle to transport passengers and do not implement MF.

Messages are generated every $25 - 35$ seconds by random cars and pedestrians. RSUs do not generate messages, but act as a communication backbone. Messages may contain pictures, video and soundbites, and are 500 KBytes to 1 MBytes in size. Suppose that memory capacity is consistent with kinds of nodes, the buffer of normal mobile phones is set to 5 MB, and smart phones, cars, RSUs, and buses have 50 MBytes buffers. We also remarked that increasing buffer size of normal phones up to 50 MBytes affects a little bit on network performance.

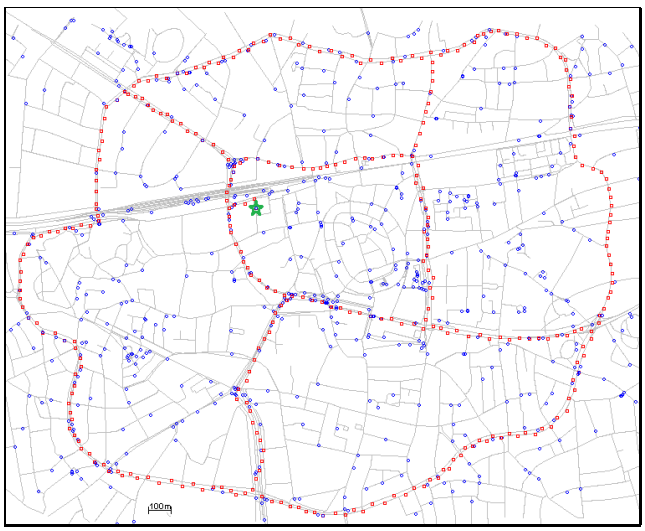


Figure 4. Inner-city of Enschede.

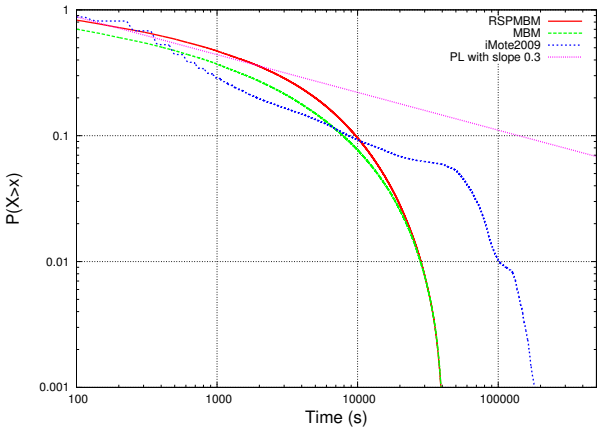


Figure 5. Inter-contact times for RSPMBM compared to the iMote trace.

B. Architecture and Model Evaluation

Figure 5 plots the complementary cumulative distribution (CCDF) of the inter-contact times. The graphs show that the inter-contact time distribution of RSPMBM has a power-law distribution with the exponent approximate 0.3 and similar to the real iMote trace [28]. This power-law distribution does contradict the exponential decay implied by previous mobility models that have been used to design routing algorithms (see [27]). Because the exponent and shape of the distribution may vary between environments, we did not configure parameters to produce the exact same exponent and shape as the iMote trace. Note that the match between the iMote trace and RSPMBM in the first two thirds of the graph. The difference in the last part of the graph is due to the longer trace (in time) of the iMote, leading to more contacts with low distribution probabilities. RSPMBM has shorter contact times due to the RSU communication backbone. In other words, nodes in our simulation environment meet one another more frequently that those in the iMote experiment.

Figure 5 also shows the inter-contact time distribution for MBM used in the Enschede City Scenario (ECS) for comparison. Surprisingly, both RSPMBM and MBM produce similar tails of distribution (exponent coefficients are about 0.3). However, the inter-contact time distribution of RSPMBM has higher probability than that of MBM. This is expected, inter-contact times usually get shorter with increasing reality [20].

C. Opportunistic Routing Algorithm Evaluation

Time-to-live (TTL) is an important variable for data dissemination, and strongly influences data delivery probability, latency, delivery speed, and delivery cost in opportunistic networks. In safety applications, emergency messages should be delivered with high probability, low latency, and high speed. Otherwise, the information might be useless after a certain period. Therefore, assigning appropriate value to TTL to drop obsolete messages probably save a lot throughput. Though TTL has a huge impact on the performance of routing protocols, it is hardly studied in existing literature. We will investigate the influence of TTL on delivery probability, latency, speed, and costs of messages.

Figure 6 shows the delivery probability of each routing algorithm as TTL in the scenarios increases from 10 to 300 minutes. In the graph two very different trends in delivery probability can be observed. DD, FC and SnW have increasing delivery probability with increasing TTL, with a highest gain in the lower TTL values. This is as one would expect. The longer the TTL of a message, the more opportunities for message transferring. Counter-intuitive is the decreasing delivery probability with increasing TTL for Epidemic and PRoPHET. This is explained as follows. Epidemic and PRoPHET are multi-copy, thus the number of relayed messages increases exponentially when TTL is long. Eventually, with a limited buffer and limited contact duration, the delivery probabilities of Epidemic and PRoPHET will dramatically suffer. This explanation is reconfirmed in Figure 9, which depicts the delivery cost for each routing protocol. We also remarked that decreasing message size or increasing buffer size lessens flooding effects on Epidemic and PRoPHET. However TTL still strongly influences their deliver probabilities.

Figure 7 plots the average latency of message delivery as TTL increases. From the graph, one can see that increasing TTL results in increasing delays of message delivery. This is as expected. Since flooding the network with messages, Epidemic scores best. Although Epidemic has the lowest delivery probability at high TTL values, when a message reaches its destination, the message will have low latency. Direct Delivery scores lowest with high latency. DD delivers messages directly to the destination. So it may take some

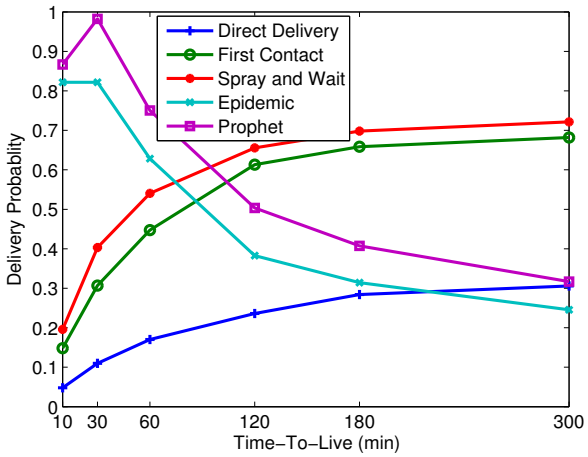


Figure 6. Message delivery probability.

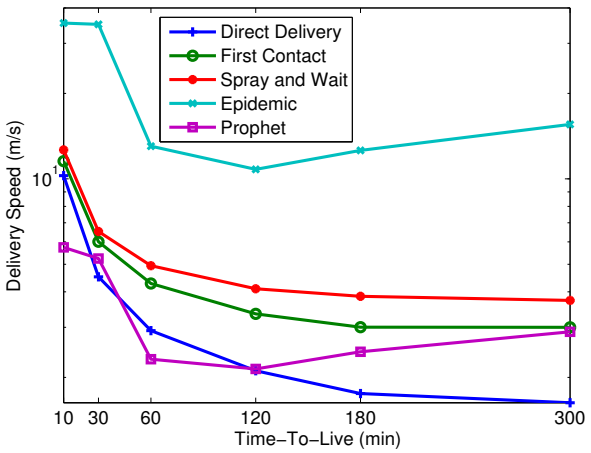


Figure 8. Average speed of message delivery.

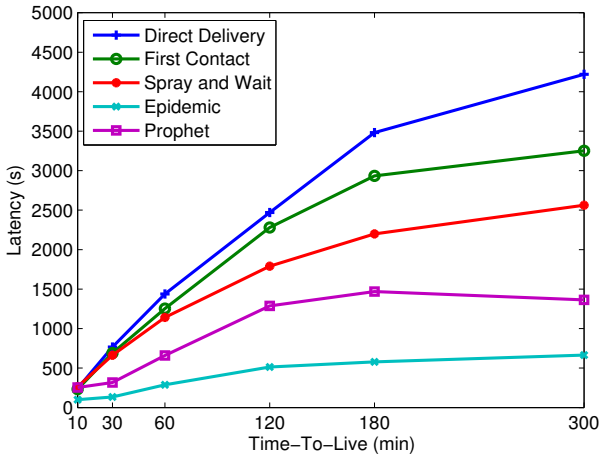


Figure 7. Average latency of message delivery.

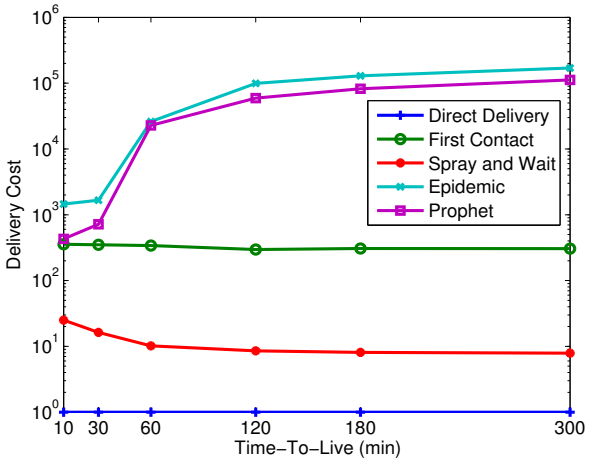


Figure 9. Delivery cost.

time for this opportunity to happen.

The speed of message delivery is depicted in Figure 8. The speed decreases sharply in the first part of the graph for all protocols and then remains almost constant. For 10-min TTL, only messages near the base station or RSUs can reach the destination. Other messages would be dropped before arriving at the base station. Increasing TTL causes more messages farther away from the base station to be delivered. This explains why the average delivery speed declines sharply. However, when TTL is greater than 60 minutes, most messages have sufficient lifetime. Therefore, increasing TTL further does not affect the delivery speed.

The delivery speed of Epidemic and PRoPHET goes up slightly when TTL is greater than 120 minutes. Due to overhead, there are fewer messages that could be delivered. Hence, the average delivery speed rises slightly again.

Epidemic has the highest delivery speed since it floods messages over the network. DD has the lowest delivery speed on account of sending messages only when mobile nodes encounter the base station.

As PRoPHET has the second lowest latency in Figure 7, one would expect it to have the second highest delivery speed. On the contrary, the graph in Figure 8 shows that PRoPHET has the lowest delivery speed when TTL is below 120 minutes. The reason lies in the fact that PRoPHET transfers messages based on the frequency of node encountering, called delivery predictability. Owing to the RSU connected network, most nodes have almost the same delivery predictability. Consequently, messages are wastefully transferred around before reaching the destination. In such way, even the average delay of a message is low, but the Euclidean distance from its source to the base station is short too. That is why the delivery speed of PRoPHET is low even though its latency is not high. This behavior

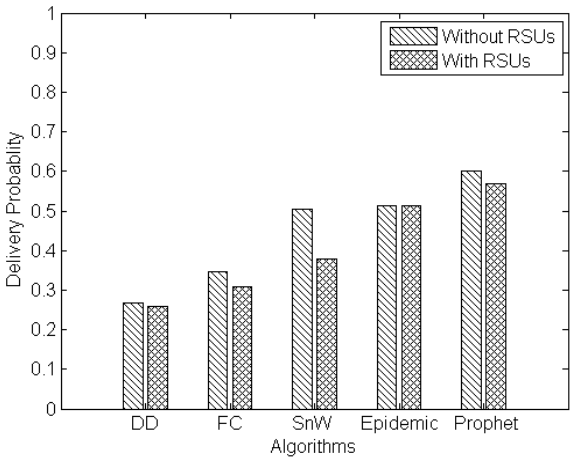


Figure 10. Delivery probability.

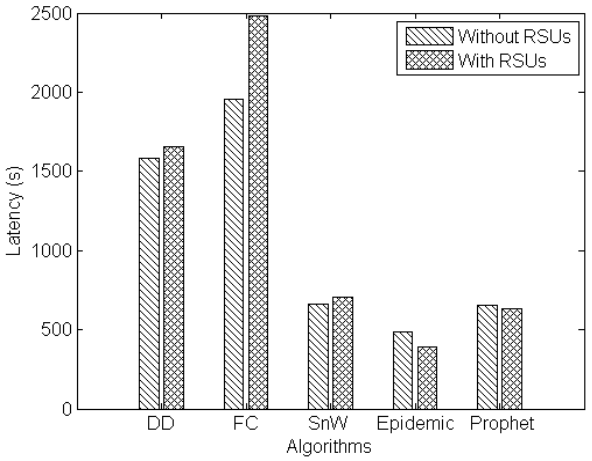


Figure 11. Latency.

also proves that delay of message delivery is not sufficient enough to evaluate quality of message delivery. However, the delivery speed of Direct Delivery even gets lower than that of PRoPHET when TTL is above 120 minutes. This is expected. When TTL increases very high, DD gives higher delay but almost constant delivery probability as in Figure 7 and 6 respectively.

Because the majority of nodes have limited power supply, the delivery costs of opportunistic routing algorithms must be taken into account. The delivery cost represents the ratio between the number of total transmissions needed over that of successfully delivered messages. Figure 9 shows that Epidemic and PRoPHET have the highest delivery cost because they maximize the opportunities of message delivery by replicating copies of messages as much as possible. DD and SnW have the least overhead, as DD has only one single copy of a message and SnW has 6 copies of messages at maximum. Clearly DD has the lowest delivery cost of all routing algorithms. The delivery costs for Epidemic and PRoPHET increase sharply with increasing TTL, but stabilize after a while. The reason is that only a limited number of messages can be transferred during the limited contact duration.

D. RSUs' Assistance Evaluation

To evaluate the advantages of RSUs in the opportunistic network, we investigate the performance of algorithms in both cases, with and without 336 RSUs. For each case, algorithms' performance will be evaluated based on four aforementioned metrics: delivery probability, delivery latency, delivery speed and delivery cost.

To make simulation more realistic, we randomly categorize messages into 5 levels of priority, from 1 (highest) to 5 (lowest). The priority means the importance or urgency of messages. Importance information, such as fire detection, should be delivered rapidly. Otherwise, it is too late, and the information is useless. It makes sense that a message with higher priority should be assigned lower time-to-live. In particular, the time-to-live value in minute is defined corresponding to message priority as in Table IV-D for our simulation.

Priority	1	2	3	4	5
Time-To-Live (min)	10	30	60	120	180

Table I
MESSAGE PRIORITY AND CORRESPONDING TIME-TO-LIVE

Such TTL assignments will significantly improve delivery performance when being combined with some buffer management. For example, Fathima and Wahidabanu [29] manage buffers by dividing messages into three sub-buffers: high priority, medium priority, and low priority. Messages with specific priority should be stored in corresponding buffer.

By intuition, we mistakenly foresaw that the present of RSUs would improve the delivery probability performed by any routing algorithm. However, simulation result in Figure 10 shows that the delivery probabilities can be worse in case of adding RSUs. This surprising conflict can be explained through studying the naive methodology of the algorithms. Since DD only transfers messages to their corresponding destinations, the existing of RSUs just scams mobile nodes with more header-list exchanges, which help nodes check the message destinations of each other. Although this wasting time is little, it still results in increasing the probability of dropped messages due to TTL, and decreases a certain number of delivered messages as well as increase the latency as shown in Figure 10 and 11, respectively. This effect is more serious for FC and SnW because both algorithms have more wasteful message

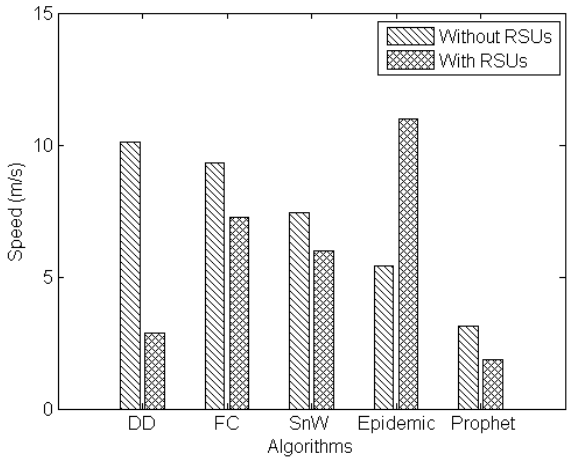


Figure 12. Delivery speed.

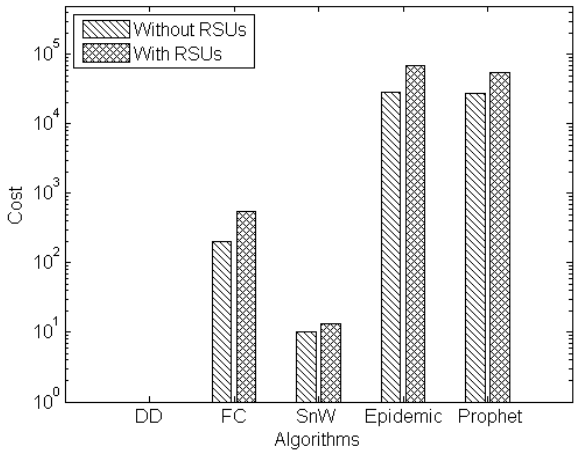


Figure 13. Delivery cost.

exchanges with enormous 336 RSUs. Remark that the hitting time of two mobile nodes is less than that of a mobile node and a RSU [14]. Therefore, both FC and SnW have lower delivery probability and higher latency in case of adding RSUs, see Figure 10 and 11.

PRoPHET also suffers from the wasting contact times with RSUs since the delivery predictability of RSUs depends on mobile nodes, not RSUs them self. In other words, transferring messages based on the delivery predictability is not suitable for stationary nodes. This explains why Prophet has the lower delivery probability in case of adding RSUs. Conversely, Epidemic has the least effect from adding RSUs in term of delivery probability and even has shorter latency since Epidemic can infinitely flood RSUs with messages. In such way, messages can be delivered faster via the connected RSU's network as shown in Figure 11.

Figure 12 shows the delivery speed of the algorithms. By comparing between Figure 11 and Figure 12, we observe that, in general, shorter latency is corresponding to higher delivery speed. However, figures show that Prophet has shorter latency but also lower speed. Again, this illustrates that latency does not totaly reflect how fast a message is delivered as we discussed. So, the extra contacts with RSUs also slows down the speed of message delivery for Prophet.

Results of delivery cost shown in Figure 13 are what we expected. Since adding RSUs leads to having more nodes in the network, there are more opportunities for transferring message copies around. Therefore, the delivery cost, which is based on the number of transmissions, increases for all algorithms, except DD because it only sends messages to destination nodes.

In conclusion, most compared algorithms cannot improve their performance in terms of delivery probability, latency, speed, and cost by adding RSUs since they are mostly designed for network of which majority are mobile nodes.

V. Conclusion and Future Work

In this paper, we have proposed a heterogeneous architecture comprising fixed infrastructure, mobile infrastructure, and mobile nodes. In addition, we propose a realistic mobility model and metrics. Several well-known opportunistic routing protocols are tested with this architecture under constraints of limited buffer size, message size, time-to-live, and unpredictable movement. Our observation shows that none of the evaluated protocols performs well with a heterogenous scenario, such as the one described in this paper. We also observe that most of the algorithms do not improve their performances when adding RSUs. Since a single simple routing algorithm does not suffice to improve the overall message delivery performance, a contribution of several algorithms should be considered:

- Road Side Units (RSU), as used in the backbone network, should not only carry received information to a central server but also disseminate information to nearby passing nodes. This communication shortcut leaves the base station out of the loop and contributes a better delivery speed and delivery cost. The Epidemic routing protocol with a flooding control mechanism is best suitable for the RSU network if delivery cost is not the most important.
- Buses, which act as data mules or message ferries, have a mobility pattern based on fixed routes and time schedules. The Message Ferry routing protocol is most appropriate.
- Pedestrians and cars are best served by stochastic and context-based schemes. However, exchanging messages between nodes that use different routing protocols is a challenge. For examples, nodes running PRoPHET fail to update the delivery predictability of nodes running Epidemic due to the unavailability of delivery predictability in Epidemic router.

We also plan to take message priority into consideration. Because designing an optimal routing protocol with a delivery probability of 100% under all conditions is difficult, prioritizing messages becomes a necessity. Message prioritization relies on the importance of information, creation time, or source location. Priorities must be defined by a specific application, for instance, public safety applications define the priority based on the source location, creation time, and seriousness of detected events. One last point of our concern is the security and privacy of information. A leading principle should be that the creator owns the data and decides how the data can be used by others. However, one may argue that in situations of emergency this principle may be overruled by authorities. This issue will be addressed in future research. Currently we are developing a heterogeneous algorithm, termed as *Unified* [30] and a testbed is planned to implement and evaluate the proposed heterogeneous algorithm.

ACKNOWLEDGMENT

This work is supported by the SenSafety project in the Dutch Commit program.

REFERENCES

[1] V.-D. Le, H. Scholten, and P. Havinga, "Towards opportunistic data dissemination in mobile phone sensor networks," in *Proc. of The Eleventh International Conference on Networks (ICN 2012)*, 2012, pp. 139–146.

[2] I. Akyildiz, W. Su, and Y. Sankarasubramaniam, "A survey on sensor networks," *IEEE Comm. Magazine*, vol. 40, pp. 102–114, 2002.

[3] Y. Wang and H. Wu, "Delay/fault-tolerant mobile sensor network (dft-msn): A new paradigm for pervasive information gathering," *IEEE Trans. Mobile Computing*, vol. 6, pp. 1021 – 1034, 2007.

[4] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc on-demand distance vector (aodv) routing," RFC Editor, United States, Tech. Rep., 2003.

[5] D. B. Johnson, D. A. Maltz, and J. Broch, "Ad hoc networking." Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001, ch. DSR: the dynamic source routing protocol for multihop wireless ad hoc networks, pp. 139–172.

[6] D. Camara, C. Bonnet, and F. Filali, "Propagation of public safety warning message: A delay tolerant approach," in *Proc. IEEE Communications Society WCNC*, 2010.

[7] A. T. Erman, A. Dilo, and P. Havinga, "A fault-tolerant data dissemination based on honeycomb architecture for mobile multi-sink wireless sensor networks," in *Proc. of the Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2010*, 2010, pp. 97–102.

[8] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Efficient routing in intermittently connected mobile networks: the multiple-copy case," *IEEE/ACM Trans. Netw.*, vol. 16, no. 1, pp. 77–90, Feb. 2008.

[9] R. Schwartz, R. Barbosa, N. Meratnia, G. Heijenk, and J. Scholten, "A directional data dissemination protocol for vehicular environments," *Computer Communications*, vol. 34, pp. 2057–2071, 2011.

[10] B. K. Polat, P. Sachdeva, M. H. Ammar, and E. W. Zegura, "Message ferries as generalized dominating sets in intermittently connected mobile networks," *Pervasive and Mobile Computing*, vol. 7, pp. 189–205, 2011.

[11] D. Borsetti, C. Casetti, C.-F. Chiasserini, M. Fiore, and J. M. Barceló-Ordinas, "Virtual data mules for data collection in road-side sensor networks," in *Proceedings of the Second International Workshop on Mobile Opportunistic Networking*, ser. MobiOpp '10. New York, NY, USA: ACM, 2010, pp. 32–40.

[12] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Department of Computer Science, Duke Univeristy, Durham, NC, Tech. Rep., 2000.

[13] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Single-copy routing in intermittently connected mobile networks," in *Proc. of Sensor and Ad Hoc Communications and Networks (SECON)*, 2004, pp. 235–244.

[14] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Spray and wait: An efficient routing scheme for intermittently connected mobile networks," in *Proc. of ACM SIGCOMM Workshop on Delay-Tolerant Networking (WDTN)*, 2005.

[15] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," in *Proc. of ACM SIGCOMM on Wireless and Delay-Tolerant Networks*, 2004.

[16] A. Lindgren and A. Droia, "Probabilistic routing protocol for intermittently connected networks," *Internet Draft draft-lindgren-dtnrg-prophet-02, Work in Progress*, 2006.

[17] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," in *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc '07. New York, NY, USA: ACM, 2007, pp. 32–40.

[18] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: social-based forwarding in delay tolerant networks," in *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc '08. New York, NY, USA: ACM, 2008, pp. 241–250.

[19] J. A. Bitsch Link, N. Viol, A. Goliath, and K. Wehrle, "Simbetage: utilizing temporal changes in social networks for pocket switched networks," in *Proceedings of the 1st ACM workshop on User-provided networking: challenges and opportunities*, ser. U-NET '09. New York, NY, USA: ACM, 2009, pp. 13–18.

[20] A. Keranen, J. Ott, and T. Karkkainen, "The one simulator for dtn protocol evaluation," in *Proc. of the 2nd International Conference on Simulation Tools and Techniques(SIMUTools)*, 2009.

[21] Newcom++, "State of the art of research on opportunistic networks, and definition of a common framework for reference models and performance metrics," Downloaded from http://www.newcom-project.eu/public-deliverables/research/DR11.1-final-1.pdf/view.

[22] R. Shah, S. Roy, S. Jain, and W. Brunette, "Data mules: modeling a three-tier architecture for sparse sensor networks," in *Sensor Network Protocols and Applications, 2003. Proceedings of the First IEEE. 2003 IEEE International Workshop on*, may 2003, pp. 30 – 41.

[23] W. Zhao, M. Ammar, and E. Zegura, "A message ferrying approach for data delivery in sparse mobile ad hoc networks," in *Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc '04. New York, NY, USA: ACM, 2004, pp. 187–198.

[24] Y. Xian, C. Huang, and J. Cobb, "Look-ahead routing and message scheduling in delay-tolerant networks," in *Proc. IEEE Conference on Local Computer Networks (LCN)*, 2010.

[25] A. Lindgren, T. Karkkainen, and J. Ott, "Simulating mobility and dtns with the one," *Journal of Communications*, 2010.

[26] F. Ekman, A. Keranen, J. Karvo, and J. Ott, "Working day movement model," in *Proc. of The 1st ACM SIGMOBILE workshop on Mobility models (MobilityModels)*, 2008.

[27] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on the design of opportunistic forwarding algorithms," in *Proc. IEEE Infocom*, 2006.

[28] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, "CRAWDAD trace cambridge/haggle/imote/infocom2006 (v. 2009-05-29)," Downloaded from http://crawdad.cs.dartmouth.edu, May 2009.

[29] G.Fathima and R. Wahidabaru, "Buffer management for preferential delivery in opportunistic delay tolerant networks," *International Journal of Wireless and Mobile Networks (IJWMN)*, vol. 3, pp. 15–28, 2011.

[30] V.-D. Le, H. Scholten, and P. Havinga, "Unified routing for data dissemination in smart city networks," in *Proc. of the 3rd International Conference on the Internet of Things (IoT2012)*, 2012.

Improving Fairness in QoS and QoE domains for Adaptive Video Streaming

Bjørn J. Villa
Department of Telematics
Norwegian Institute of Science and Technology
Trondheim, Norway
bjorn.villa@item.ntnu.no

Poul E. Heegaard
Department of Telematics
Norwegian Institute of Science and Technology
Trondheim, Norway
poul.heegaard@item.ntnu.no

Abstract - **This paper presents an enhancement to a category of Adaptive Video Streaming solutions aimed at improving both Quality of Service (QoS) and Quality of Experience (QoE). The specific solution used as baseline for the work is the Smooth Streaming framework from Microsoft. The presented enhancement relates to the rate adaptation scheme used, and suggests applying a stochastic or fixed/unique setting of the rate adjustment intervals rather than the default fixed/equal approach. The main novelty of the paper is the simultaneous study of both network oriented fairness in the QoS domain and perception based fairness from the QoE domain, when introducing the suggested mechanism. The method used for this study is by means of simulations, measurements and numerical optimization. Perception based fairness is suggested as an objective QoE metric which, requires no reference to original content. The results show that the suggested enhancement has potential of improving fairness in the QoS domain, while maintaining perception based fairness in the QoE domain.**

Keywords - Adaptive Video Streaming; Fairness; QoE/QoS.

I. INTRODUCTION

Solutions for Adaptive Video Streaming are part of the more general concept of ABR (Adaptive Bit Rate) streaming [1] which, covers any content type. The implementation of ABR streaming for video varies between different vendors, and among the more successful one today is the Microsoft Smooth Streaming framework [2]. In general, the different implementations use undisclosed and proprietary functions, even across interfaces between client and server. The latter is addressed by the new MPEG DASH standard [3].

The basic behavior of adaptive video streaming solutions is that the client continuously performs a measurement and estimation of available resources in order to decide which, quality level to request. The relevant resource from the network side is the available capacity along the path between the server and client. Based on this, at certain intervals the client decides to either go up or down in quality level or remain at the current level. The levels are predefined and communicated to the client by the server at session startup. The changes in quality levels are normally done in an incremental approach, rather than by larger jumps in rate level. The rationale behind this is the objective to provide a smooth watching experience for the user. However, it may also be related to the CPU monitoring done by the client, as this is a key resource required. It may be the case that even if the network can provide you with a much higher rate level, the CPU on the device being used would not be able to process it. During the initial phase of an adaptive streaming session the potential requests of change in rate level are more frequent than later on when operating in a more steady-state phase. To some extent this is a rather aggressive behavior from a single client which, may have undesirable inter-stream impacts. At the same time, in order to give the user a good first impression and make him want to continue using the service it is desirable to reach a high quality as soon as possible.

Among the strongest drivers for commercial use of ABR based services on the Internet are Over-The-Top content providers. These are providers which, rely on the best effort Internet service as transport towards their customers. Therefore, technologies aiming at making services survive almost any network state are of great interest. In addition to focus on the network based QoS dimensions of services and involved networks, there is also a growing interest in the QoE dimension [4]. The latter should be considered as not only a richer definition of quality, but also more focused towards who decides whether something is good or bad, i.e., the end user. The evolution of successful services on Internet indicates that the focus on QoE for Over-The-Top providers is a good strategy.

A. Problem Statement

The concept of Adaptive Video Streaming is very promising. However, as more and more services are adopting this concept the success brings new challenges. The first challenge with effects visible to the end users is how well these services behave when they compete for a shared resource, such as a home broadband access. With a strong dominance of video based service on the Internet this issue is important to address. As each client operates independently of each other, it has no understanding of the traffic it competes with. Different clients consider each other as just background traffic. This leads to unpredictable behavior of each session. The focus of this paper is to study a method for improving QoS/QoE fairness among competing streams in a home network environment.

B. Research Approach

The method investigated in this paper to address the problem at hand is to apply specific changes in the algorithm used by each ABR client controlling the adaptive

behavior. The specific change suggested is related to the rate adjustment interval used [2]. The effect of changing the duration of the rate adjustment interval from an equal T duration to either a random or per session unique duration is presented and analyzed.

The ABR solution used as reference point for the work is the one from Microsoft (Smooth Streaming). However, the key principles would still apply to other solutions based on similar principles.

C. Paper Outline

The structure of this paper is as follows. Section II presents related work; Section III provides an overview of methodology and metrics; Section IV describes the simulation model; Section V presents simulation results; Section VI presents the measurement setup together with results; In Section VII the simulation results and measurements results are summarized and compared; In Section VII an analytical view of the methods studied are given; Section IX provides the conclusions and an outline of future work.

II. Related Work

It has been shown in [5] that competing adaptive streams can cause unpredictable performance for each stream, both in terms of oscillations and ability to achieve fairness in terms of bandwidth sharing. The experimental results presented give clear indication on that competing ABR clients cause degraded and unpredictable performance. Apart from this paper, the topic at hand does not seem to have been addressed by the academic research community to the extent it deserves.

In another paper [6], the authors have investigated how well adaptive streaming performs when being subject to variable available bandwidth in general. Their findings were that the adaptive streams are performing quite well in this type of scenario except for some transient behavior. These findings do not contradict the findings in [5] as the type of background traffic used do not have the adaptive behavior itself, but is rather controlled by the basic TCP mechanisms.

Rate-control algorithms for TCP streaming in general and selected bandwidth estimation algorithms are described in [7]. This work is relevant to any TCP based application delivering a video stream.

In some of our own previous work we have described and analyzed how competing adaptive streams can be controlled using a knowledge based bandwidth broker in the home gateway [8] [9]. We have also developed a testbed for performing experimental verification of methods studied [10] which has been used for collecting the measurement used in this paper.

III. Methodology and metrics

In this section, we introduce the relevant performance metrics together with motivation for the chosen focus. Thereafter, some candidate methods on how to improve the performance metrics are given, and finally, the specific method subject for study is presented.

A. Flow Based Performance Metrics

For transport flows it is common [11] to focus on the following metrics in order to assess their performance: inter-flow fairness, stability and convergence time. This in addition to the general QoS metrics: bandwidth, packet loss, delay and jitter. The same metrics can be applied to adaptive video streams as they by definition also are flows with similar concerns. The analysis of these metrics can be done from a strict network oriented perspective (QoS), but to some extent also bridged over to a user perception domain (QoE). When focusing on the inter-flow fairness metric this is traditionally analyzed [12] using, e.g., the Jain's fairness index [13], the product measure [14] or Epsilon-fairness [15] for flows with equal resource requirements. For flows with different resource requirement, the Max-Min fairness [16], proportional fairness [17] or minimum potential delay fairness [18] approaches are commonly seen. Real life adaptive video streams would typically belong to the last category.

Max-Min fairness: The objective of max-min fairness is to maximize the smallest throughput rate among the flows. When this is met, the next-smallest throughput rate must be as large as possible, and so on. Max-min fairness can also be explained by considering it as a progressive filling algorithm, where all flows start at zero and grow at the same pace until the link is full. With this approach the max-min fairness gives priority to the smallest flows. The least demanding flows always have the best chance of getting access to all the resources it needs.

Proportional fairness: The original definition of proportional fairness comes from economic disciplines [17] for the purpose of charging. The original definition is used in the relevant RFC [12] but it does not come across as very constructive for the purpose of analyzing fairness in single resource (e.g., bandwidth) sharing among flows. In this context more recent definitions and interpretations are more suitable [19]. The principle of this would be that a resource allocation is considered proportional fair if it is made to the flow which, has the highest ratio between potential maximum resource consumption and its average resource consumption so far. A further simplification would be to use the current resource usage (if greater than 0) instead of the average in the ratio calculation. The same ratio numbers for each flow could then be used to give a view on the current system fairness by comparing them. If they are all equal the system could be stated as proportionally fair.

Minimum potential delay fairness: The idea behind minimum potential delay fairness is based on the assumption that the involved flows are generated by applications transferring files of certain sizes. A relevant bandwidth sharing objective would be to minimize the time needed to complete those transfers. However, this does not

apply to an adaptive streaming scenario and is therefore not discussed any further.

B. Perception Based Performance Metrics

There is a wide range of metrics which, influence how satisfied an end user is with a service such as e.g., video streaming. Many of these are not related to network aspects, and therefore difficult to influence by means in this domain. However, one of the perceived performance metrics which, could be correlated with network aspect is the notion of perceived fairness. It is then of great interest to try and find methods of influencing this in a positive manner.

Looking at fairness from an end user perception, research from the social science and psychology domain [20] states that this is closely related to what is called 'Social Justice'. In this context a queuing system or any other resource allocation mechanism would be considered as a 'Social System'. It has further been found that users react negatively to any system behavior which, gives better service to other users, unless justification is provided. Such system behavior is considered un-fair, i.e., in violation with the social justice of the system as the end users considers it as discrimination.

The end user notion of system discrimination has been suggested by [21] as an important measure of perceived service quality, and more specifically the perceived fairness is stated to be closely related to the discrimination frequency. It should be noted that analyzing this type of end user perceived discriminations has a challenge in terms of handling the false positive and false negative cases.

Applying the concept of discrimination to competing adaptive streams, it would be related to situations where end user expectations are not met during steady state periods and also negative changes in service delivery during more transient periods. In other words, whatever makes the end user think that he is being discriminated due to other users in the system, will lead to reduced perceived service quality.

In order to use this type of perceived end user discrimination as a measure for how well the algorithm which, controls the adaptive streams are performing, a clear definition regarding what end users are considering as discrimination is required. This could, e.g., be periods with session rate below some threshold, any change in session rate to a lower level or the session rate change frequency.

C. QoS and QoE Fairness

Based on the overview given in the previous sections for both flow based and perception based performance metrics, the following definitions are presented for the fairness metrics subject for study in this paper.

In the QoS domain, we use proportional fairness as the key metric while in the QoE domain we use perceived fairness, defined as follows.

Proportional Fairness - The difference between the worst and best performing streaming sessions in terms of average rate achieved during the session lifetime divided by session max rate.

Perceived Fairness – The difference between the worst and best performing streaming sessions in terms of average number of rate reductions (i.e. discrimination events) per minute.

Following this, the main focus is put on differences in performance for the worst and best performing sessions. However, the absolute values for both achieved session rate and session quality level reductions are of course also relevant when evaluating the proposed methods.

D. Methods for Improving Performance

There are several things that one could try to incorporate into the adaptive algorithms controlling the ABR service in order to make them perform better in a multi-stream scenario.

The selected performance metrics to be studied are proportional fairness and perceived fairness metric as described. Whether it is possible to improve both these fairness metrics at the same time will be an important part of the results. We consider the following approaches as interesting to consider in this domain.

Randomization or unique time intervals: The equal rate adjustment intervals (*T*) used by each adaptive stream while in steady-state may be a contributing factor to inaccurate estimations of available bandwidth and thereby oscillating behavior. An alternative to fixed intervals would be to randomize them by using a per-session stochastic parameter (within certain reasonable bounds) or assigning each session a unique value. By doing so the available bandwidth estimation methods may become more accurate.

Back-off periods: Whenever a service is reducing its rate level due to observed congestion it may try to increase again after the same amount of time (*T*). In addition to the previous described randomization/unique approach to this interval, one could also consider introducing a back-off period. This would imply that after a service has reduced its rate level, it enters a back-off period of a certain duration during which, no increase is allowed.

Threshold based behavior: Rather than using the same intervals of potential rate changes all the time, one could introduce a threshold for when it operates more or less aggressive. This threshold could be the mean available rate level for a specific session, or even a smoothed average value for the actual achieved level. This concept is applied with success in more recent TCP versions for the purpose of optimizing performance.

The method chosen for this study is according to the first approach described, i.e., using a random or unique interval between each potential rate change. This would represent a different approach than the default method used in Smooth Streaming from Microsoft [2].

As baseline for the simulations, the default interval *T=2s* has been used. Then as alternatives, both a stochastic distribution and per session fixed unique distribution has

been implemented. For the stochastic approach the Uniform distribution was chosen with parameters [1.6, 2.4]s. For the fixed unique approach, the sessions were spread on the following value set [1.6, 1.8, 2.0, 2.2, 2.4]s.

IV. SIMULATION MODEL

As the adaptive streaming solutions of today are highly proprietary, the details concerning their implementation are not disclosed. Due to this, there will always be some degree of uncertainty concerning their internal functions.

The simulation model is based on our earlier work [1] but has been somewhat simplified in order to allow for comparison with experimental results.

A. Assumptions

One of the key functions of an ABR client is the method used for determining whether to go up or down in rate level during times of varying available bandwidth. From studying live traffic it does not seem as if the clients use additional network probing beyond the actual information obtained through download of video segments. Further on, in the likely absence of a per stream traffic shaper at the server side (for scalability and performance reasons), it will give a traffic pattern for each stream which, typically contains a sequence of burst and idle periods. The measured burst period rate is then higher than the actual stream rate level. Also, it is likely that there will be sub-periods within the burst periods where per packet rate is close to the total available bandwidth. As such, the client can probably obtain a rather accurate indication of maximum available bandwidth by just looking at minimum observed inter-arrival time of packets of known size belonging to the same stream.

However, not all streams will have interleaved burst periods so there is a good chance for each stream to overestimate the potential for additional bandwidth. There is a wide range of bandwidth estimation methods and a few of these are described in [22], but again - as the details of the adaptive streaming solutions are not disclosed we will not discuss this part any further. Independent of which, method being used, there will be some degree of uncertainty which, contributes to variable performance. Further on, we assume the following to be true for the ABR sessions to be studied

- No stream coordination at server side
- No involvement from mechanisms in the network between the client and server
- All clients operate independently and do not communicate
- All clients are well behaved in the sense that they follow the same scheme
- At each defined stream rate level there are no variations due to i.e., picture dynamics
- All clients access the same stream on the server side

B. Session Type and Schedule

The ABR sessions used in the simulator are based on profiles observed in commercial services. The quality levels defined are {0, 350, 500, 1000, 1500, 2000, 3000, 4000, 5000} Kbps. All sessions are of the same type. The sessions are initiated by 5 different users and start time scheduling are done according to the stochastic distributed parameter t_a – Uniform [0, 2000]ms. This gives that all sessions start during the first 2 seconds.

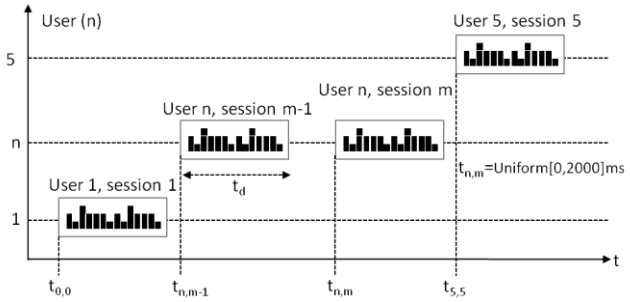


Figure 1. Session scheduling per user

During one simulation run, each user executes a total of 100 sessions sequentially. Time for starting the next session (m) for specific user (n) is noted $t_{n,m}$ (cf. Figure 1). The duration of each session t_d is deterministic and set to 25 minutes.

C. Rate Adaptation Algorithm

The model for rate adaptation per session is based on periodic estimation of available bandwidth $A_s(t)$ and calculation of a smoothed average $SA_s(t)$.

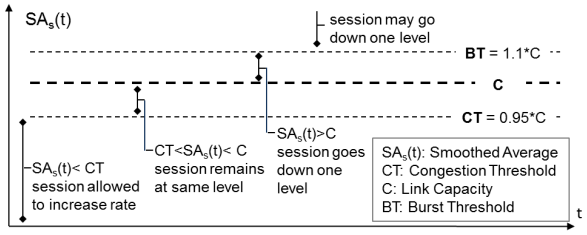


Figure 2. Thresholds for smoothed average

This smoothed average (cf. Figure 2) is compared to a congestion threshold (*CT*), the link capacity (*C*) and a burst threshold (*BT*) in order to trigger a rate adjustment.

Whenever the sum of requested rates from sessions is above the burst threshold (*BT*), the next session which, calculates $SA_s(t)$ will be forced down, independent of the value of $SA_s(t)$. This function is implemented in the simulator in order to incorporate the somewhat unpredictable behavior during times of heavy congestion.

The calculation of smoothed average $SA_s(t)$ is based on [5], and is expressed in (1). The parameter δ gives the weighting of the estimated available bandwidth for the two periods included in the calculation.

$$SA_s(t) = \delta\, A_s(t_{i-1}) + (1-\delta) A_s(t_i) \qquad (1)$$

The available bandwidth estimation function used in the simulations is based on the assumption that sessions running at high rates are able to make more accurate estimations than those running at lower rates. An abstraction of the function itself is made by a number of n bandwidth samples $C_{i,j}$ (cf. Figure 3)

A specific session is then given access to a number of these samples according to its current rate level, and then it will use this as basis for its estimation. A high rate gives a high number of samples available, and then, also, a higher degree of accuracy.

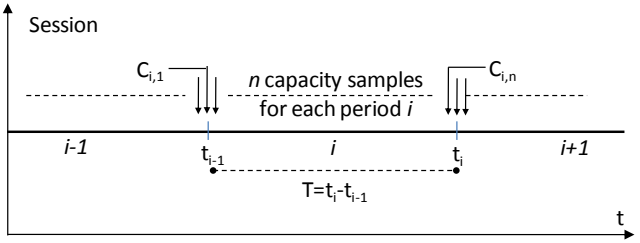


Figure 3. Capacity samples per period

The number of samples $x_{s,i}$ available to a specific session *s* for period *i* is given by its ratio between current rate $R_s(t_i)$ and max rate $R_s max$, multiplied by *n* as per (2).

$$x_{s,i} = n^{R_s(t_i)}/_{R_s max} \qquad (2)$$

In the simulations, the value of n was set to 20 and $R_s max$ was according to the session definition 5000Kbps. The available bandwidth estimated $A_s(t)$ for period *i* is then given by the following (3).

$$A_s(t_i) = \sum_{l=1}^{x_i} C_{i,l}/_{x_{s,i}} \qquad (3)$$

By combination with the expression for $SA_s(t)$ it gives the following expression (4).

$$SA_s(t) = \delta \sum_{l=1}^{x_{i-1}} C_{i,l}/_{x_{s,i-1}} + (1-\delta) \sum_{l=1}^{x_i} C_{i,l}/_{x_{s,i}} \qquad (4)$$

The value of δ was set to 0.8 as per [5], thus giving most weight to the available bandwidth estimation from the previous period.

D. Simulation Tool

The simulator was built using the process oriented Simula [23] programming language and the Discrete Event Modeling On Simula (DEMOS) context class [24].

This programming language is considered as one of the first object oriented programming languages, and remains a strong tool for performing simulations.

V. Simulation Results

The simulation results are presented for different capacity levels on the access link. The chosen capacities are 10, 12.5, 15, 17.5 and 20Mbps. At all these capacity levels there would be congestion as the sum of the maximum quality level requested for the 5 competing sessions is 25Mbps. The simulations were also run for capacity levels between those given above, but for the sake of clarity these details are left out as they did not change the conclusions.

Simulation session results are sorted and then grouped according to the studied metrics, giving a clear view on performance ranging from the worst to the best performer.

The characterization is done by looking at the distributional properties location (mean), spread (mid 50% values) and high/low 25% results. For this purpose the box and whisker plots are used as they give a compact view of all these properties.

A. Proportional Fairness

As defined, proportional fairness is calculated by the achieved session average rate per user, divided by session max – and then a comparison of these values are done for the competing sessions/users. The results from the simulations give 100 independent samples for this metric.

Improvements in proportional fairness are then recognized as reduced difference between the worst and best performing sessions. The results are presented in Figure 4, Figure 5 and Figure 6 showing both the mean values and the spread of the metric sample distributions.

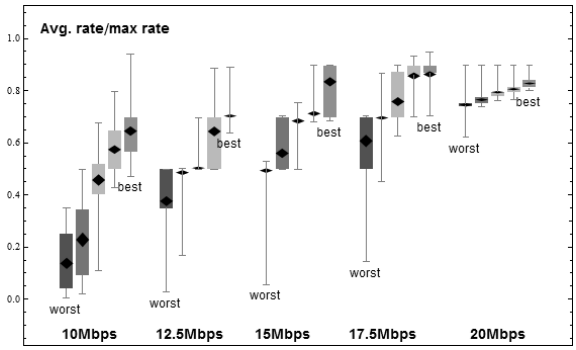


Figure 4. Proportional Fairness, Equal T, Simulations

The results shown in Figure 4 illustrates that there is a significant challenge in terms of proportional fairness when

using the default equal T approach for all access capacity levels except for at the highest level (20Mbps).

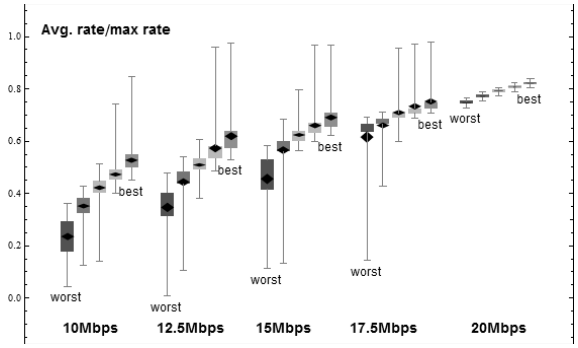


Figure 5. Proportional Fairness, Unique T, Simulations

By careful study of the results shown in Figure 5 for the unique T approach one can see that the difference between the worst and best performing sessions are reduced, and thereby an improved proportional fairness.

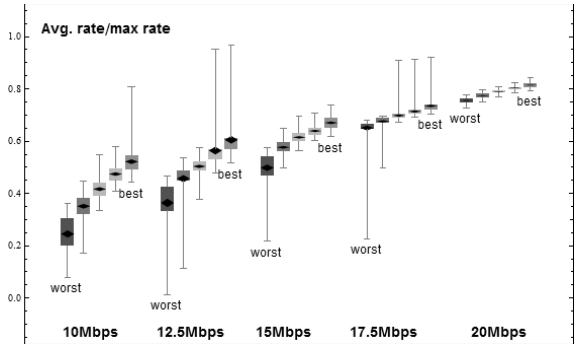


Figure 6. Proportional Fairness, Random T, Simulations

The same effect as for the unique T approach is also visible for the random T approach as shown in Figure 6. For both approaches it is also worth noticing the reduced spread of observations as indicated by the mid 50% values.

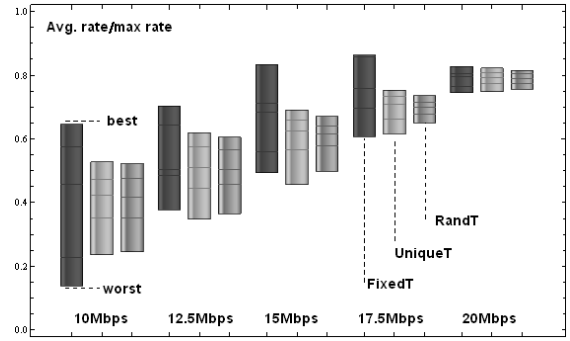


Figure 7. Summary Proportional Fairness, Simulations

In the summary view of the simulations results as shown in Figure 7 the effect of both random T and unique T methods are quite clear. The differences between the competing sessions become smaller, thus we can state that the simulations give reason to believe that the methods studied give improvements in terms of proportional fairness.

B. Perceived Fairness

As follows by our definition of perceived fairness a small difference between sessions in terms of number of rate reductions per minute is good. The rationale behind this would be an assumption of that different users have insight into the performance of other sessions. In addition, the absolute value is of course also important. A low metric value is good.

The results shown in Figure 8 give a clear indication on that the simulator model is quite aggressive in terms of how often it allows each stream change its quality level. The level of 15 reductions / minute is likely to represents the model maximum. This follows by T intervals of 2 sec, and our presentation of reductions / minute only.

The results for perceived fairness using the equal T approach are quite poor in the sense that the absolute values are at maximum level for the three mid capacity levels. However, it should be noted that the simulator model contains some simplifications and assumptions which may not be accurate enough in this domain.

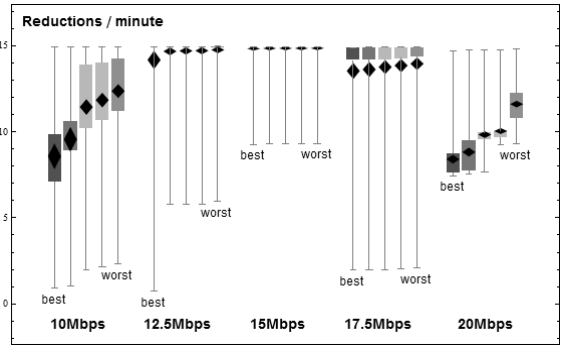


Figure 8. Perceived Fairness, Equal T, Simulations

The results shown Figure 9 for the unique T approach illustrates that the reductions per minute are reduced, but at the same time it introduces a stronger difference between sessions.

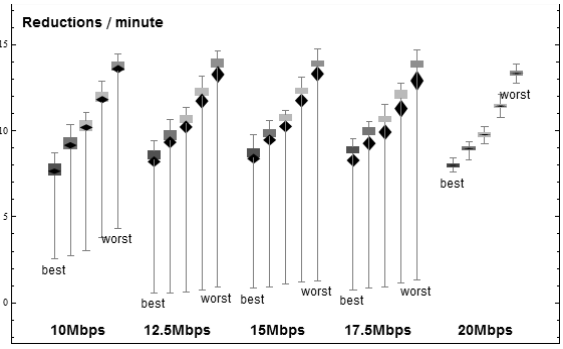


Figure 9. Perceived Fairness, Unique T, Simulations

The same effect as for the unique T approach is also visible for the random T approach as shown Figure 10. Except for the higher spread at 20Mbpss capacity levels, the results are quite similar.

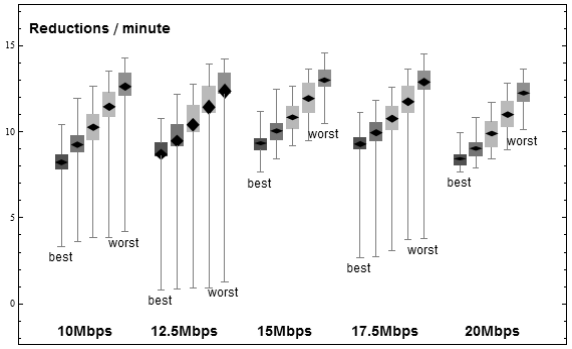


Figure 10. Perceived Fairness, Random T, Simulations

The summary view of perceived fairness is presented in Figure 11. It illustrates both the actual values for the best/worst performers and the difference between them. Results are presented for the default equal T, unique T and random T methods for all access capacity levels. These results alone do not give reason to believe that the investigated method (unique T and random T) give an improved perceived fairness.

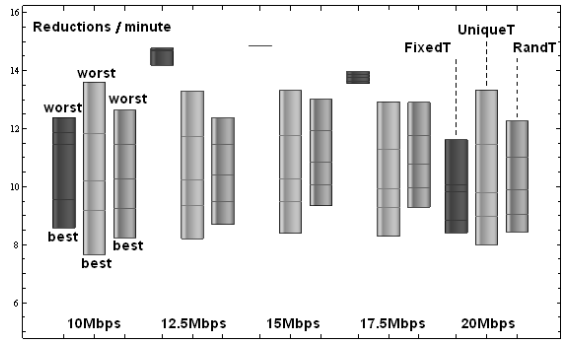


Figure 11. Summary Perceived Fairness, Simulations

VI. MEASUREMENTS

In order to perform the required measurements a testbed was established in a controlled environment including all required componenst [10].

As illustrated in Figure 12 the 5 clients are located behind a shared access with a certain capacity towards a Microsoft Smooth Streaming service. This scenario matches the one which was built into the simulator as described in section IV.

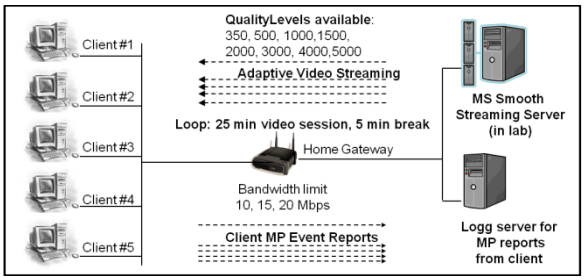


Figure 12. Measurement setup

The clients used were separate PC's with identical HW and SW and set to access the same adaptive HTTP video stream from the lab server. Controlled by scripts on each PC the clients were run in a loop with intervals of 25 minute active streaming and then 5 minute break.

For each scenario studied the loop was set to give 100 interval repetitions. An earlier developed tool for event reporting [25] from each client (Monitor Plane event reports) was used in order to record interesting events on a per session basis and allow for effective post processing.

The measurements results for proportional fairness and perceived fairness are given in the following sections using the same presentation form as for the simulations.

A. *Proportional Fairness*

The results shown in Figure 13 illustrate that there is a problem with regards to proportional fairness when using the default equal T approach for all access capacity levels. The problem is smallest at the highest level (20Mbps), which matches the earlier presented simulation results (cf. Figure 4).

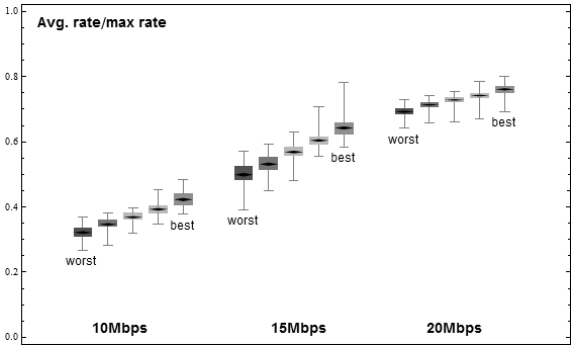


Figure 13. Proportional Fairness, Equal T, Measurements

By studying the results shown in Figure 14 for the unique T approach, a noticeable reduced difference between the worst and best performing sessions are seen. This again, is in line with the corresponding simulation results (cf. Figure 5) indicating improved proportional fairness.

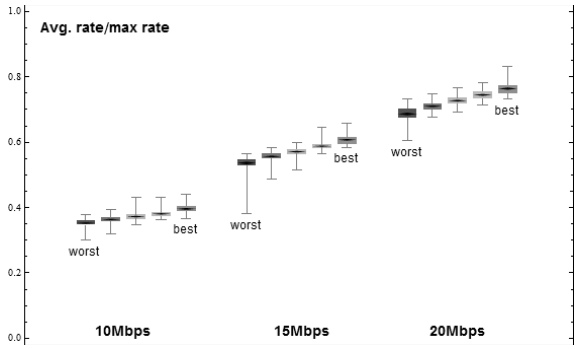


Figure 14. Proportional fairness, Unique T, Measurements

A similar positive effect as for the unique T approach is also visible for the random T approach (cf. Figure 15).

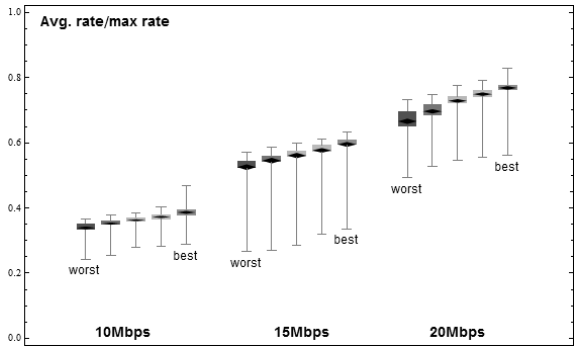


Figure 15. Proportional Fairness, Random T, Measurements

In the summary view of the measurements results as shown in Figure 16 the positive effect of both random T and unique T methods are quite clear, except for at the highest access capacity level (20Mbps). These findings are much in line with the finding from the simulations (cf. Figure 7).

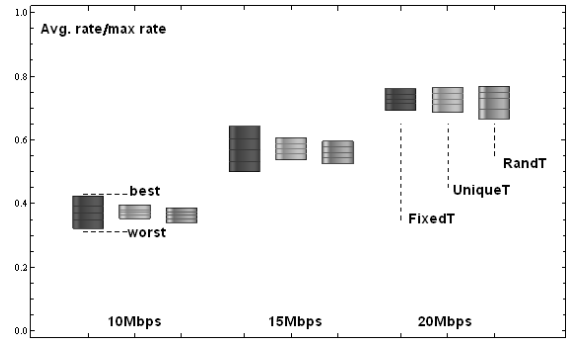


Figure 16. Summary Proportional Fairness, Measurements

It should be noted that measurements were not performed for all the capacity levels which were used in the simulations. The main reason for this is the amount of time required for performing measurements versus time required for simulations.

B. *Perceived Fairness*

The first thing which is noticed when looking at the measurements result for perceived fairness in Figure 17 is that the levels observed are much lower than those collected during simulations (cf. Figure 8). Thereafter, one can see that there is a clear difference between the best and worst performing sessions but the absolute values are rather low.

Therefore, based on these findings we can only state that there is a pure theoretical challenge with perceived fairness. Whether actual users will feel discriminated or get a poor user experience due to quality fluctuations at these levels is not evident.

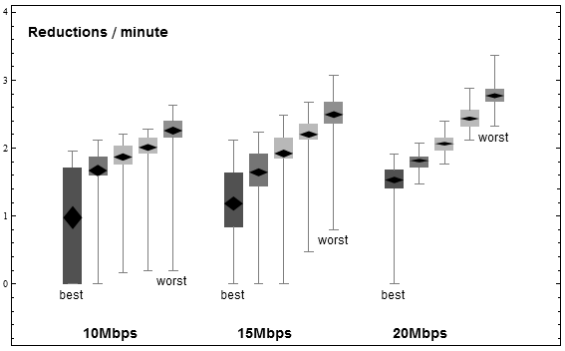


Figure 17. Perceived Fairness, Equal T, Measurements

The results shown Figure 18 for the unique T approach illustrates that the spread in the observations are reduced (mid 50% observations), but the mean value levels remain in the same regions as for the default equal T approach.

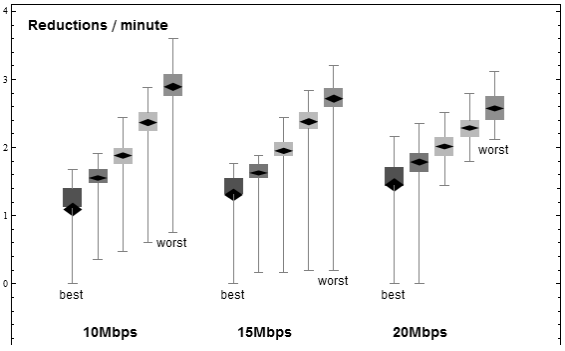


Figure 18. Perceived Fairness, Unique T, Measurements

For the random T approach as illustrated in Figure 19 we see an increase in spread for the observations at the two lowest access capacity levels, making the results in this regard almost similar to the default equal T approach. The exception is the results for 20Mbps access where a quite clear positive effect is seen with regards to difference between the worst and best performing session.

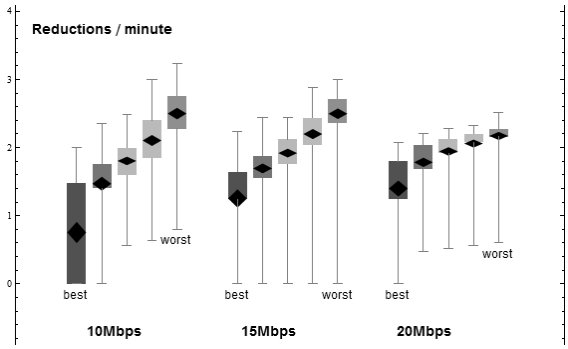


Figure 19. Perceived Fairness, Random T, Measurements

The summary view of perceived fairness based on measurements is presented in Figure 20. As can be seen, the results do not give reason to state an improvement in terms of perceived fairness when implementing either the unique T or random T methods.

These findings are in line with the simulation results, although there is a major difference in the absolute levels.

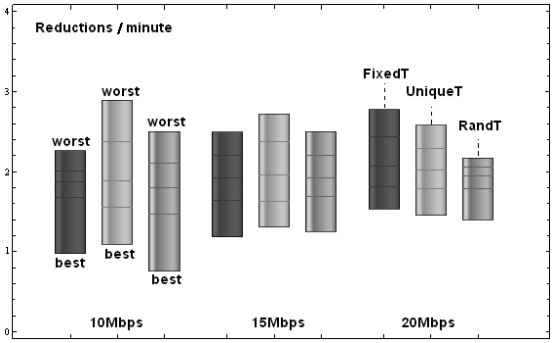


Figure 20. Summary Perceived Fairness, Measurements

VII. COMPARING SIMULATIONS AND MEASUREMENTS

The results from simulations and measurements differ in absolute values for both proportional fairness and perceived fairness. Keeping in mind that any simulation is based on a model and not the real system itself, this does not come as surprise. However, the important thing is to highlight the effect of introducing the suggested methods (random T, unique T) and see if there are similarities in this regard in both the simulation and measurement domains.

Looking at the combined results for proportional fairness given in Figure 21 we see that a similar effect is present in both domains. There is a clear positive effect of introducing either the random T or unique T method.

Both the simulation results and measurements results show a very strong positive effect for most access capacity levels, except for at the highest level (20Mbps) where the effect is close to neutral.

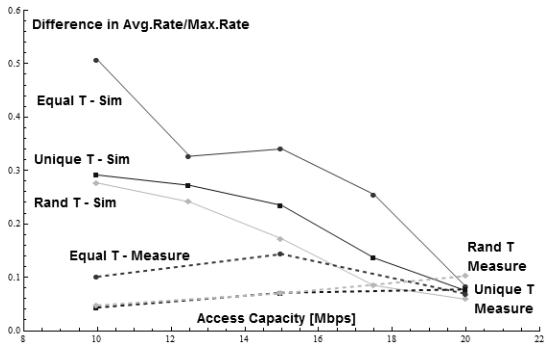


Figure 21. Proportional Fairness, Measurements and Simulations

For the perceived fairness results as shown in Figure 22 the simulation part indicates a strong improvement for the suggested methods. However, as the absolute values are so high (close to assumed maximum) the credibility of these results is weakened. The rate adaptation algorithm implemented in the simulator is probably too aggressive compared to the real life implementations.

The measurement results for perceived fairness are neutral viewed alone, but when combined with the proportional fairness results one can say it is positive that improvements in the pure QoS domain does not come at the expense of degraded performance in the QoE domain.

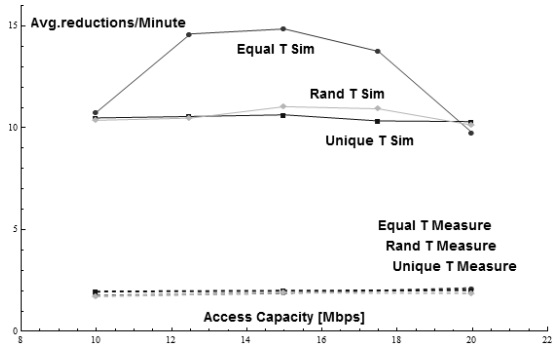


Figure 22. Perceived Fairness, Measurements and Simulations

In summary, the simulations together with the measurements gives a strong indication of that the suggested methods have a potential real life value in terms of improving proportional fairness.

The differences between using a random T value or a per session unique T value does not give basis for saying which is better. However, from an implementation point of view the random approach clearly has its challenges as the video content requires encoding according to these intervals. In light of this, the approach of using per session unique T values is the preferred one.

VIII. ANALYSIS

The somewhat intuitive explanation to why changes could be expected when introducing either a random T or unique T rate adjustment interval is that some of the

negative effects of an equal adjustment interval as illustrated in Figure 23 are reduced. In the case of equal periods, each session would get the same periodic view on the link utilization, always missing or including some other traffic. This gives a certain degree of error in the available bandwidth estimation functions.

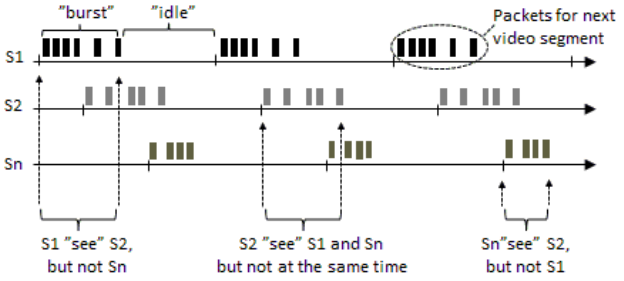


Figure 23. Problem with equal estimation periods

Each session estimates available bandwidth only during its burst periods. Although not explicitly stated in system documentation [2] this has been verified in the measurement testbed [10]. As part of this work the absence of any active network probing was verified. Therefore, whatever notion of available bandwidth each client uses as basis for its rate adaptation it must be based on information collected during the periods where it receives video segments (burst periods). This means that in order to get an accurate estimation it is beneficial for each client to have overlapping burst periods with as many other sessions as possible.

A. Burst Period Duration

The duration of the burst period for a specific session depends on both its current rate level and the rate adjustment interval. The dependency of the rate level follows from the obvious relation to data volume to be transferred per time unit for a specific rate level, while the dependency of rate adjustment interval follows from the requirement to maintain the same average amount of data received over time.

At the beginning of each interval the client requests the next video fragment for a specific rate level, with duration equal to its rate adjustment interval. This is illustrated in Figure 24 where two sessions running at the same rate level, but with different rate adjustment intervals have different burst period durations.

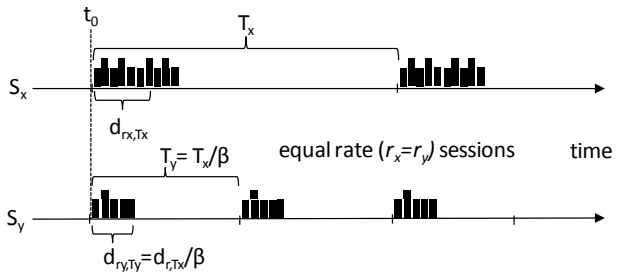


Figure 24. Equal rate sessions with different burst periods

Any two sessions (S_x, S_y) running at the same rate level, will have a relation between their burst period durations expressed by the parameter β. This parameter is given by the following expression (5).

$$\beta = {d_{rx,Tx}}/{d_{ry,Ty}} = {T_x}/{T_y}, \qquad for\ r_x = r_y \tag{5}$$

Using this relationship, we can express (6) the burst period duration $d_{ri,Ti}$ for any session S_i as a function of its rate adjustment interval T_i and a reference burst period duration $d_{ri,T}$.

$$d_{ri,Ti} = d_{ri,T}({T_i}/{T}) \tag{6}$$

The values for $d_{ri,T}$ can presumably be calculated based on information about the codec used for the specific media stream inside each sessions, together with assumption on per session server side capacity. Alternatively one could make measurements on a specific system and establish a $d_{ri,T}$ matrix for all valid values of r_i and the reference T value.

However, if we assume that the server side capacity is not a limitation, and that it will always try to burst with a certain bitrate C_{burst} we can also express the burst period duration $d_{ri,Ti}$ as follows (7).

$$d_{ri,Ti} = \left({r_i\,T}/{C_{burst}}\right)\left({T_i}/{T}\right) = \left({r_i T_i}/{C_{burst}}\right) \tag{7}$$

The maximum value for C_{burst} is natural to think of as the access capacity for the user group / home network, as this is normally the end-to-end bottleneck. However, it is likely that the actual C_{burst} is related to the maximum rate for the specific service.

B. Probability for Burst Period Overlap

For T_i values according to a uniform distribution, the probability $P_{i,r,t}$ for a session i at rate level r to be in its burst period at time t will be according to the following expression (8).

$$P_{i,r,t} = {d_{r,Ti}}/{T_i} = \frac{d_{r,T}\left({T_i}/{T}\right)}{T_i} = {d_{r,T}}/{T} \tag{8}$$

From this, we see that all sessions at a specific rate level has the same probability of being in its burst period at time t. We can then express the probability that all n sessions are in their burst period at time t as follows (9).

$$P_{all\ burst,t} = \left({d_{r_1,T}}/{T}\right)^{c_1} \left({d_{r_m,T}}/{T}\right)^{c_m} \tag{9}$$

The parameter c_m represents the number of sessions at rate level r_m and the sum of all c_m values equals n. From this we see that the probability of any session to see all other sessions during its burst period depends on the session rate level mix, and this probability increases when more sessions are running at high rate levels.

Further on, we recognize that the probability for that a session *i* has an overlap with each of the other sessions sometimes during its burst period T_i is the integral of $P_{all\ burst,t}$ over the period [0, T_i] which, is easily expressed as the constant $P_{all\ burst,t}$ multiplied by T_i.

We then let a specific session mix be described by the vector $R_{mix}=\{r_1,\ldots,r_n\}$, whereas r_i represents the rate level for session *i*. Also, for a specific session *i* let A_i be the group of sessions which, has overlapping burst periods with session *i* at a specific time t_0, and B_i be the group of sessions for which, it did not have an overlap. In the situation where all sessions have the same rate adjustment interval duration T_i, the probability of that session *i* has an overlapping burst period with any of the sessions in group B_i at time t_0+T_i is zero. This leads to that while R_{mix} remains unchanged, the view a specific session has of the total traffic will not change. The system state for session *i* in terms of burst period overlap with other sessions is independent of the state at t_0 and also *t* in general.

In the case where T_i is not equal for all sessions, but instead are chosen according to some stochastic distribution – the group of sessions which, overlap the burst period of session *i* at t_0+T_i is not independent of the state at t_0. If we let C_i denote the sub-group of sessions from B_i which, has overlapping burst periods with session *i* at time t_0+T_i, it can be shown that there is a deterministic relationship between A_i, B_i and C_i.

If we then remember the assumed use of a smoothed average function we see the benefit of this potential additional burst period overlaps in subsequent periods.

C. Dynamics in Burst Period Overlap

When the starting times for each session and their respective rate adjustment intervals (T_i) are considered stochastic processes, the sessions will combine in time in different ways. In order to define the deterministic relationship between overlapping burst periods during subsequent intervals, we need to analyze scenarios where sessions with different rate levels and different rate adjustment interval are combined.

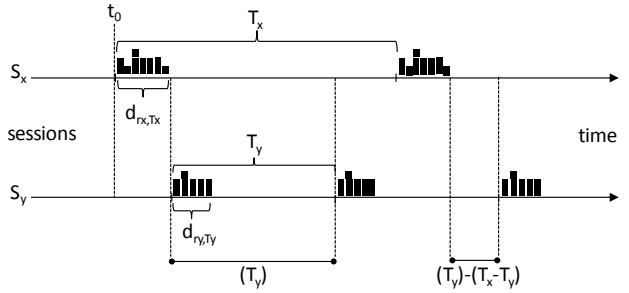


Figure 25. Session S_y staring after S_x (T_y<T_x)

The first scenario (*a*) to be studied is the one where two sessions (S_x, S_y) with different T_i values (T_x, T_y) are active at the same time. We assume $T_x > T_y$ and that S_y starts immediately after the burst period of S_x finishes as illustrated in Figure 25.

For the two sessions (S_x, S_y) there will be shift in phase between them as a function of time which, makes them have a full or partial burst period overlap at some time. The question is then how many rounds it will take for S_x to see S_y and vice versa. It can be shown that we can express the number of rounds for S_x before it has an overlapping burst period with S_y as follows (10).

$$\begin{aligned} \mathrm{N}_{a,x\to y} &= 1+\left\lceil \frac{T_y}{T_x-T_y} \right\rceil \\ \text{when } \tfrac{T_x}{2} < T_y < (T_x - d_{rx,Tx} - d_{ry,Ty}) & \\ \mathrm{N}_{a,x\to y} &= 2 \\ \text{when } (T_x - d_{rx,Tx} - d_{ry,Ty}) < T_y < T_x & \end{aligned} \quad (10)$$

In the same way, we can express the number of rounds for S_y before the same overlap of burst period with S_x takes place (11).

$$\begin{aligned} \mathrm{N}_{a,y\to x} &= 1+\left\lceil \frac{T_x}{T_x-T_y} \right\rceil \\ \text{when } \tfrac{T_x}{2} < T_y < \left(T_x - d_{rx,Tx} - d_{ry,Ty}\right) & \\ \mathrm{N}_{a,y\to x} &= 2 \\ \text{when } \left(T_x - d_{rx,Tx} - d_{ry,Ty}\right) < T_y < T_x & \end{aligned} \quad (11)$$

The next scenario (*b*) to be studied is where the sessions (S_x, S_y) are running with different T_i values (T_x, T_y) but now S_y finishes its burst period before S_x (cf. Figure 26).

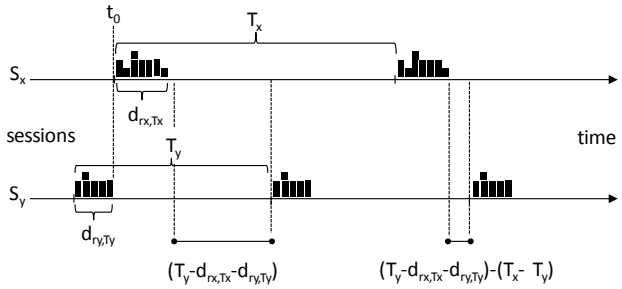


Figure 26. Session S_x staring after S_y (T_y<T_x)

The number of rounds it takes for S_x to see S_y is expressed as follows (12).

$$\begin{aligned} \mathrm{N}_{b,x\to y} &= 1+\left\lceil \frac{T_y-dx-dy}{T_x-T_y} \right\rceil \\ \text{when } \tfrac{T_x}{2} < T_y < \left(T_x - d_{r,Ty}\right) & \\ \mathrm{N}_{b,x\to y} &= 2 \\ \text{when } (T_x - d_{r,Ty}) < T_y < T_x & \end{aligned} \quad (12)$$

The number of rounds it takes for S_y to see S_x is expressed as follows (13).

$$N_{b,y\rightarrow x} = 1 + \left\lceil \frac{T_x - dx - dy}{T_x - T_y} \right\rceil$$
$$\text{when } \frac{T_x}{2} < T_y < (T_x - d_{r,Ty}) \qquad (13)$$
$$N_{b,y\rightarrow x} = 2$$
$$\text{when } (T_x - d_{r,Ty}) < T_y < T_x$$

It should be noted that for both scenarios there is a special case where $N_{a,y\text{-}x}/N_{b,y\text{-}x}$ and $N_{a,x\text{-}y}/N_{b,x\text{-}y}$ are always 2, i.e., two sessions which, did not have overlapping burst periods at t_0 is guaranteed to have overlapped during the next period for S_x and S_y. For a smoothed average function operating over two periods this is desirable, i.e., whatever it does not see in the first period it is guaranteed to see in the next.

D. Optimization Problem

The expressions for $N_{y\text{-}x}$ and $N_{x\text{-}y}$ contain many variables. These variables are the rate adjustment intervals T_i and the burst period durations $d_{ri,Ti}$ for all sessions. The latter are calculated based on the session rates r_x and r_y and C_{burst} as defined in Section V. These expressions can be used as input to a constrained optimization problem and analyzed as such in order to find maximum and minimum values.

As the starting point for this optimization problem we can focus on the worst case scenario, that would be the number of rounds for S_y before it has an overlap with S_x ($N_{a,y\text{-}x}/N_{b,y\text{-}x}$), which, will always be higher than the number of rounds for S_x before this has an overlap with S_y.

We also see that $N_{a,y\text{-}x}$ will always be greater than $N_{b,y\text{-}x}$ since $T_x > T_y$. This gives us only one expression to analyze for the worst case scenario as follows (14).

Maximize: $N_{a,y\rightarrow x}$

where

$$N_{a,y\rightarrow x} = \begin{cases} 1 + \left\lceil \frac{T_x}{T_x - T_y} \right\rceil, if\ T_y < \left(T_x - d_{rx,Tx} - d_{ry,Ty}\right) \\ 2, if\ \left(T_x - d_{rx,Tx} - d_{ry,Ty}\right) < T_y < T_x \end{cases} \qquad (14)$$

subject to:

$$1.6 < T_y, T_x < 2.4 \text{ and } T_x/2 < T_y$$
$$r_x, r_y \in \{350,500,1000,1500,2000,3000,4000,5000\}$$
$$d_{r_x,T_x} = {r_x T_x}/{C_{burst}}$$
$$d_{r_y,T_y} = {r_y T_y}/{C_{burst}}$$

The above maximization can then be done for different values of C_{burst}. In the simulations and measurements the access capacities used were between 10 and 20Mbps and the maximum session rate was 5Mbps. Based on measurements of real traffic we can see that the C_{burst} is lower than the actual access speed and therefore values of respectively 5Mbps, 7.5Mbps and 10Mbps were used for C_{burst}.

For the two different alternatives of choosing values for T_i used in the simulations and measurements, both the random T and unique T approaches are possible to work with in the optimization context. However, as the unique T approach will be a special case (subset) of the random T, we only present results for the random T approach.

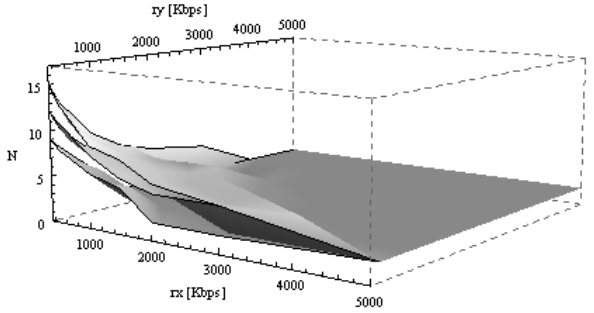


Figure 27. Maximum $N_{a,y\text{-}x}$ for different burst bitrates

The result from solving the optimization problem is shown in Figure 27. The three different burst bitrates (C_{burst}) give surfaces which, are plotted, whereas the highest capacity gives the highest values for $N_{a,y\text{-}x}$.

We see that in many cases we get an overlap already in the second round, and thereby we improve the basis for the available bandwidth estimation algorithm. This analysis then strengthens the findings in both the simulations and measurements.

In order to improve the available bandwidth estimations further one may consider the well known PASTA principle [26] from queuing theory which, states that a Poisson based Arrival process See Time Averages. This implies that the bandwidth probing should take place not only during the burst periods, but as a process taking samples throughout the whole rate adjustment period. However, as this implies some degree of active probing it would potentially have some other undesirable effects.

IX. Conclusions and Future Work

The results show that there is a significant potential of improving proportional fairness as defined while maintaining perceived fairness for adaptive streams of the category studied. The positive effect of the suggested enhancement to the rate adaptation scheme, i.e., using a random or unique duration of rate adjustment intervals rather than the default equal value is supported by simulation results, measurements and also rationalized by the theoretical analysis. The findings differ to some extent from those in our previous work [1], but at the same time we now have a more refined and accurate view of the methods

studied. The added value of results from measurements has been significant.

The results also illustrate that when studying the performance of adaptive streaming solutions, it is not enough to only focus on the network centric QoS domain. A change in this domain does not necessary lead to a corresponding change in the QoE domain, and vice versa.

As future work in this field it is planned to further study objective and no-reference based QoE metrics which, is possible to correlate over to the QoS and network domain. It is also planned to study various available bandwidth algorithms with regard to their real-time capabilities and thereby suitability for adaptive video streaming.

X. ACKNOWLEDGEMENTS

The reported work is done as part of the Road to media-aware user-Dependant self-aDaptive NETWORKS - R2D2 project. This project is funded by The Research Council of Norway.

The work has also been actively supported by TV2, the leading commercial TV broadcaster in Norway. TV2 is among the pioneers in providing a full commercial TV offering over the Internet based on ABR technology.

REFERENCES

[1] B. Villa and P. Heegaard, "Improving perceived fairness and QoE for adaptive video streams," *ICNS 2012*, March 2012. In Proceedings of ICNS 2012: The Eighth International Conference on Networking and Services, Thinkmind 2012, ISBN 978-1-61208-186-1, March 2012, pp. 149-158.

[2] A. Zambelli, "IIS Smooth Streaming Technical Overview," Tech. Rep., March 2009.

[3] ISO, *Dynamic adaptive streaming over HTTP (DASH). ISO/IEC FCD 23001-6*, International Organization for Standardization Std. ISO/IEC FCD 23001-6, 2011.

[4] E. Areizaga, L. Perez, C. Verikoukis, N. Zorba, E. Jacob, and P. Odling, "A road to media-aware user-dependent self-adaptive networks," in *Proc. IEEE Int. Symp. BMSB '09*, 2009, pp. 1–6.

[5] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over http," in *ACM Multimedia Systems (MMSys)*, 2011.

[6] L. De Cicco and S. Mascolo, "An experimental investigation of the akamai adaptive video streaming," in *Proceedings of the 6th international conference on HCI in work and learning, life and leisure: workgroup human-computer interaction and usability engineering*, ser. USAB'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 447–464.

[7] R. Kuschnig, I. Kofler, and H. Hellwagner, "An evaluation of tcp-based rate-control algorithms for adaptive internet streaming of h.264/svc," in *MMSys '10*. New York, NY, USA: ACM, 2010, pp. 157–168.

[8] B. J. Villa and P. E. Heegaard, "Monitoring and control of QoE in media streams using the click software router," in *NIK2010, Norway. ISBN 978-82-519-2702-4.*, vol. 1, November 2010, pp. 24–33.

[9] B. Villa and P. Heegaard, "Towards knowledge-driven QoE optimization in home gateways," *ICNS 2011*, May 2011, Thinkmind, ISBN 978-1-61208-133-5, pp. 252-256.

[10] B. J. Villa, P. E. Heegaard, and A. Instefjord, "Improving fairness for adaptive http video streaming," in *EUNICE* 2012, Springer 2012, ISBN 978-3-642-32807-7, pp. 183–193.

[11] S. Bhatti, M. Bateman, and D. Miras, "Revisiting inter-flow fairness," in *Broadband Communications, Networks and Systems, 2008. BROADNETS 2008. 5th International Conference on*, sept. 2008, pp. 585 –592.

[12] S. Floyd, "Metrics for the Evaluation of Congestion Control Mechanisms," RFC 5166 (Informational), Internet Engineering Task Force, Mar. 2008.

[13] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared systems," DEC, Tech. Rep., 1984.

[14] D. Mitra and J. B. Seery, "Dynamic adaptive windows for high speed data networks with multiple paths and propagation delays," *Computer Networks and ISDN Systems*, vol. 25, no. 6, pp. 663 – 679, 1993, high Speed Networks.

[15] Y. Zhang, S.-R. Kang, and D. Loguinov, "Delayed stability and performance of distributed congestion control," in *Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications*, ser. SIGCOMM '04. New York, NY, USA: ACM, 2004, pp. 307–318.

[16] Z. Cao and E. Zegura, "Utility max-min: an application-oriented bandwidth allocation scheme," in *INFOCOM '99*, vol. 2, mar 1999, pp. 793 –801 vol.2.

[17] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, 1997.

[18] S. Kunniyur and R. Srikant, "End-to-end congestion control schemes: utility functions, random losses and ecn marks," *IEEE/ACM Trans. Netw.*, vol. 11, pp. 689–702, October 2003.

[19] A. Jdidi and T. Chahed, "Flow-level performance of proportional fairness with hierarchical modulation in ofdma-based networks," *Comput. Netw.*, vol. 55, pp. 1784–1793, June 2011.

[20] H. Levy, B. Avi-Itzhak, and D. Raz, "Network performance engineering," D. D. Kouvatsos, Ed. Berlin, Heidelberg: Springer-Verlag, 2011, ch. Principles of fairness quantification in queueing systems, pp. 284–300.

[21] W. Sandmann, "Quantitative fairness for assessing perceived service quality in queues," *Journal Operational Research*, pp. 1–34, April 2011.

[22] R. Prasad, C. Dovrolis, M. Murray, and K. Claffy, "Bandwidth estimation: metrics, measurement techniques, and tools," *Network, IEEE*, vol. 17, no. 6, pp. 27 – 35, nov.-dec. 2003.

[23] R. J. Pooley, *An Introduction to Programming in Simula*. Blackwell Scientific Publications. ISBN: 0632014229, 1987.

[24] G. Birtwistle, *Demos - A system for Discrete Event Modelling on Simula*, G. Birtwistle, Ed. School of Computer Science, University of Sheffeld,, July 1997.

[25] B. J. Villa and P. E. Heegaard, "A monitor plane component for adaptive video streaming," in *NIK2011, Norway. ISSN 1892-0713*, vol. 1, November 2011, pp. 145–154.

[26] R. W. Wolff, "Poisson Arrivals See Time Averages," *Operations Research*, vol. 30, no. 2, pp. 223–231, 1982.

Mitigating Some Security Attacks in MPLS-VPN Model "C"

Shankar Raman*, Balaji Venkat†, and Gaurav Raina†
India-UK Advanced Technology Centre of Excellence in Next Generation Networks
*Department of Computer Science and Engineering, †Department of Electrical Engineering
Indian Institute of Technology Madras, Chennai 600 036, India
Email: mjsraman@cse.iitm.ac.in, balajivenkat299@gmail.com, gaurav@ee.iitm.ac.in

***Abstract*—In certain models of inter-provider Multi-Protocol Label Switching (MPLS) based Virtual Private Networks (VPNs), spoofing and replay attacks against VPN sites are two key concerns. MPLS VPN model "C" can scale well with respect to maintenance of routing state when compared with models "A" and "B". But this deployment model is not favoured due to the aforementioned security concerns in the data-plane. The inner labels associated with VPN sites are not encrypted during data transmission. Therefore it is possible for an attacker to spoof or replay data packets to a specific VPN site. We propose a label-hopping technique which uses a set of randomised labels and a method for hopping amongst these labels to address these type of attacks. To reduce the computation time complexity for such algorithms, we propose the use of Timing over Internet Protocol connection and Transfer of Clock (TicToc) based Precision Time Protocol. Simulations show that by using the TicToc protocol, along with the label-hopping technique, we can mitigate spoofing and replay attacks at line-rate. As we address key security and performance concerns, we make a plausible case for the deployment of MPLS based VPN inter-provider model "C".**

***Index Terms*—MPLS; VPN; Model "C"; Label-hopping; Spoofing attack; Replay attack.**

I. Introduction

Mitigating spoofing and replay attacks in Multi-Protocol Label Switching - Virtual Private Networks (MPLS-VPNs) is a key concern [1]. MPLS [2] technology uses fixed size labels to forward data packets between routers. Specific customer services (for example, Layer 3 (L3)-VPNs based on Border Gateway Protocol (BGP) extensions), can be deployed by stacking the labels. BGP-based MPLS L3-VPN services are provided either on a single Internet Service Provider (ISP) core or across multiple ISP cores. The latter cases are known as inter-provider MPLS VPNs, which are broadly categorised and referred to as models "A", "B" and "C" [3].

Model "A" uses back-to-back VPN Routing and Forwarding (VRF) connections between Autonomous System Border Routers (ASBRs). Model "B" uses exterior BGP (eBGP) redistribution of labelled VPN Internet Protocol version 4 (IPv4) routes from Autonomous Systems (AS) to neighbouring AS. Model "C" uses multi-hop Multi-Protocol (MP)-eBGP redistribution of labelled VPN IPv4 routes and eBGP redistribution of IPv4 routes from an AS to a neighbouring AS. Model "C" is scalable for maintaining routing states and hence preferred for deployment in the Internet [4]. Security issues in MPLS, especially MPLS-based VPNs, continue to attract attention [5].

The security of model "A" matches the single-AS standard proposed in [6]. Model "B" can be secured on the control-plane, but on the data-plane the validity of the outer-most label (Label Distribution or Resource Reservation Protocol label) is not checked. This weakness could be exploited to inject crafted packets from inside an MPLS network. A solution for this problem is proposed in [4]. Model "C" can be secured on the control-plane but has a security weakness on the data-plane. The ASBRs do not have any VPN information and hence the inner-most label cannot be validated. In this case, the solution used for model "B" cannot be applied. An attacker can exploit this weakness to send unidirectional packets into the VPN sites connected to the other AS. Therefore, Internet Service Providers (ISPs) using model "C" must either trust each other or not deploy it [7]. A simple solution to this problem is to filter all IP traffic with the exception of the required eBGP peering between the ASBRs, thereby preventing a large number of potential IP traffic-related attacks. However, controlling labelled packets is difficult. In model "C", there are at least two labels for each packet: the Provider Edge (PE) label, which defines the Label Switched Path (LSP) to the egress PE, and the VPN label, which defines the VPN associated with the packet on the PE.

Control-plane security issue in model "C" can be resolved by using IPSec [8]. The authors propose an IPSec encryption technique for securing the PE of the network. The authors also highlight that the processing capacity could be over-burdened. Further, if IPSec is used in the data-plane then configuring and maintaining key associations could be difficult. If an attacker is located at the core of the network, or in the network between the providers that constitute an inter-provider MPLS VPN, then spoofing is possible. The vulnerability of MPLS against spoofing attacks and the impact on performance of IPSec has been discussed in [9]. If the inner labels that identify packets going towards a L3-VPN site are spoofed, then sensitive information related to services available within the organisational servers can be compromised.

The algorithm previously proposed by us to mitigate spoofing attacks is an $O(N)$ algorithm, where N represents the payload size chosen for hashing [1]. However, using payload to obtain the hash value can encourage replay attacks on a VPN site. It should be noted that the labels used in the

label-hopping algorithm are valid only for a certain period of time. An attacker could resend a valid data packet within this time period. The label-hopping algorithm accepts such packets. Such an attack reduces the network performance as redundant data packets get processed repeatedly. A simple way to solve this problem is to include a sequence number with every packet, but this increases the payload size. Therefore label-hopping with hashing based on payload cannot be used to provide protection against replay attacks.

In this paper, we expand the work presented in [1] in the following ways:

1) We use Timing over IP Connection and Transfer of Clock (TicToc) to achieve label synchronisation and hence mitigate replay as well as spoofing attacks.
2) We show that use of TicToc, hashing and pseudo-random number generators to mitigate replay attacks leads to a constant time ($O(1)$) computational time complexity increase to the algorithm that mitigates spoofing attacks.

Additionally, we show that the computational time complexity of the label-hopping algorithm can be reduced from $O(N)$ to $O(1)$ by using time-based synchronisation techniques like Network Time Protocol (NTP) or TicToc. Such methods will be useful in a real time data transfer scenario in MPLS VPNs as they incur very low processing overhead. The advantage of the proposed scheme is that it can be used wherever MP-eBGP multi-hop scenarios arise. We also show that the proposed method can reduce the burden on Deep Packet Inspection Engines (DPIEs), but can contribute towards more processing time for ISP's billing schemes. As far as we know, no ISP has implemented MPLS VPN model "C". Large scale deployment of this model has been avoided due to security concerns. The methods proposed in this paper make a case for the potential deployment of MPLS VPN model "C' by ISPs.

The rest of the paper is organised as follows. In Section II, we discuss the pre-requisites of the proposed scheme. Section III reviews the label-hopping technique. In Section IV, we present a method by which the computational time complexity can be reduced using TicToc. In Section V, we present algorithms that protect model "C" against spoofing and replay attacks. In Section VI, we present our simulation results. Some of the implementation issues are discussed in Section VII. In Section VIII, we present the impact of label-hopping scheme on two applications; namely deep packet inspection engine and ISP's billing. Section IX outlines our contributions, and highlights some avenues for further work.

II. Pre-requisites for the label-hopping scheme

We briefly review the network topology for model "C", the PE configuration and the control-plane exchanges needed for the proposed scheme.

A. MPLS VPN model "C"

The reference MPLS-eBGP based VPN network for model "C" as described in [10] is shown in Figure 1, along with the control-plane exchanges. A legend for Figure 1 is given in Table I. The near-end PE (PE_{ne}) and far-end PE (PE_{fa}) are connected through the inter-provider MPLS based core network. The VPN connectivity is established through a set of routers from different AS and their ASBRs. In the VPN, MP-eBGP updates are exchanged for a set of Forward Equivalence Classes (FECs). These FECs, which have to be protected, originate from the prefixes behind PE_{ne} in a VPN site or a set of VPN sites.

B. PE configuration

Various configurations are needed in the PEs inside the Autonomous Systems (AS) to implement the label-hopping scheme. These are listed below:

1) A set of "m" algorithms that generate collision-free labels (universal hashing algorithms) are implemented in the PEs. Each algorithm is mapped to an index $A = (a_1, a_2, \cdots a_m), m \geq 1$. Ordering of the algorithms must be the same in the PEs. If the PEs used are from different vendors then a standardised set of algorithms must be used.
2) The bit-selection pattern is used by the PE. This helps in determining the bits chosen for generating the additional label. This additional label plays a role in avoiding collision in the hash values.
3) PE_{ne} is configured for a FEC or a set of FECs represented by an aggregate label (per VRF label). For each FEC or a set of FECs, a set of valid labels used for hopping, $K = (k_1, k_2, k_3, \cdots k_n), n > 1$ and, $k_i \neq k_j$ if $i \neq j$, is configured in PE_{ne}. This helps in selective application of the schemes for the FECs. In the case of bi-directional security, the roles of the PEs are reversed.

C. Control and data-plane flow

Initially, set K and the bit-selection pattern used by the PEs are exchanged securely over the control-plane. Optionally an index from A, representing a hash-algorithm, could also be exchanged. We propose that only the index is exchanged between the PEs, as it enhances the security for two reasons. First, the algorithm itself is masked from the attacker. Second, the algorithm can be changed frequently, and it would be difficult for the attacker to identify the final mapping that generates the label to be used for a packet. Figure 1 depicts this unidirectional exchange from PE_{ne} to PE_{fa}.

Once the secure control-plane exchanges are completed, we apply the label-hopping technique. PE_{fa} forwards the labelled traffic towards PE_{ne} through the intermediate routers using the label-stacking technique (Figure 2). The stacked labels along with the payload are transferred between the AS and ASBRs before they reach PE_{ne}. Using the label-hopping algorithm PE_{ne} verifies the integrity of labels. Upon validation, PE_{ne} uses the label information to forward the packets to the appropriate VPN service instance or site. This data-plane exchange from PE_{fa} and PE_{ne} is depicted in Figure 3. A legend for Figure 3 is given in Table I. Figure 3 also shows how the labels for the packets are specified when the data packets flow from CE2 to CE1. In the figure, the L3

header network address is $172.16.10.1$ whose gateway is CE1. We now present the label-hopping scheme.

Abbreviation	Description
AS	Autonomous Systems
ASBR	Autonomous System Border Router
CE	Customer Edge Routers
LDP: L1-L4	Label Distribution Protocol with link labels
NH	Next Hop
PE	Provider Edge Routers
POP, V1	Label between AS1 and PE_{ne}
VPN	Virtual Private Network

TABLE I: Legend for Figures 1 and 3

III. LABEL-HOPPING TECHNIQUE

Once a data packet destined to the PE_{ne} arrives at the PE_{fa} a selected number of bytes from the payload is chosen as input to the hashing algorithm. The resulting hash-digest is used to obtain the first label for the packet. The agreed bit-selection pattern is then applied on the hash-digest to determine an additional label, which is then concatenated with the first label. Once PE_{ne} receives these packets it verifies both the labels.

The implementation steps for the control-plane at the PE_{ne} and PE_{fa} are given by Algorithm 1 and Algorithm 2. The implementation steps for the data-plane at the PE_{fa} and PE_{ne} are given by Algorithm 3 and Algorithm 4.

A brief explanation of these algorithm follows:

Algorithm 1 exchanges four attributes, namely

1) the acceptable Forward Equivalence Classes (FECs),
2) valid and acceptable labels for each of the FECs,
3) the pointer or instance to the hash algorithm, and
4) the bit selection pattern to be used, with the PE_{fa} using a secure control-plane exchange.

Step 3 of Algorithm 1 assumes that the function *CP-SendPacket()* sends secure encrypted data packet to PE_{fa}.

Algorithm 1 Control-plane PE_{ne} algorithm

Require: FEC[] Forward Equivalence Classes, K[] valid labels, A[i] hash algorithm instance, I[] the bit-selection pattern chosen for the inner label.

```
1: Begin
2: packet = makepacket(FEC,K, A[i], I);
3: CP-SendPacket(PE_fa, MP-eBGP, packet);
4: End
```

Algorithm 2 receives the secure packet, decrypts it and then fills up its tables by extracting the FECs and the label mapping of the FECs. It then selects the hash algorithm based on the instance or the pointer passed by the PE_{ne}. These are done in steps $3-7$. We assume that both the PEs implement the same hash algorithms corresponding to the pointers or instances that are passed. Note that this is pre-configured in the routers. The PE_{fa} also gets to know the valid bit selection pattern that is acceptable for the PE_{ne} in step 8.

Algorithm 2 Control-plane PE_{fa} algorithm

Require: None

```
1: Begin
2: packet = CP-ReceivePacket(PE_ne); // from PE_ne
3: FEC[] = ExtractFEC(packet); // extract FECs
4: K[] = ExtractLabels(packet); // extract the labels
5: selectHashAlgorithm(A[i]); // hash algorithm to use
6: RecordValues(FEC); // information for PE_fa
7: RecordValues(K); // information on the keys
8: RecordValues(I); // bit-selection pattern to be used
9: End
```

Algorithm 3 describes the processing that occurs before the data packets are sent from PE_{fa}. Steps $3-6$ in the algorithm checks whether the label-hopping algorithm is enabled for the FEC. If it is not enabled, the algorithm will proceed to exchange data packets without label-hopping. If the label-hopping algorithm is enabled for the FEC, then the hash-digest of the packet, as well as the first and additional labels are generated at steps $7-9$. The data packet is then encapsulated with the labels and sent to the PE_{ne}.

Algorithm 3 Data-plane PE_{fa} algorithm

Require: None

```
1: Begin
2: packet = DP-ReceivePacket(Interface);
3: match = CheckFEC(packet); // Is the algorithm enabled?
4: if match == 0 then
5:    return; // algorithm not enabled.
6: end if
7: hash-digest = calculateHash(A[i],packet);
8: first-label = hash-digest % |K|;
9: addl-label = process(hash-digest,I)
10: DP-SendPacket(PE_ne, first-label, addl-label, packet);
11: End
```

Algorithm 4 receives the encapsulated packet from PE_{ne}. It then determines whether the FECs deploy the label-hopping scheme; see steps $3-6$. In steps $7-11$, the algorithm extracts the labels from the packet and calculates the hash-digest for the packet as well as the inner and additional labels. It compares the calculated values with the extracted values of the labels; see steps $12-17$. If a match exists on the labels sent by PE_{fa}, then the packet is considered to be valid. The data packets are passed to the CE after removing the labels that match.

Figure 4 gives a modified version of a sequence diagram for all the four algorithms discussed in this section. This diagram also partially shows the calls executed by the PEs in the control and data-planes.

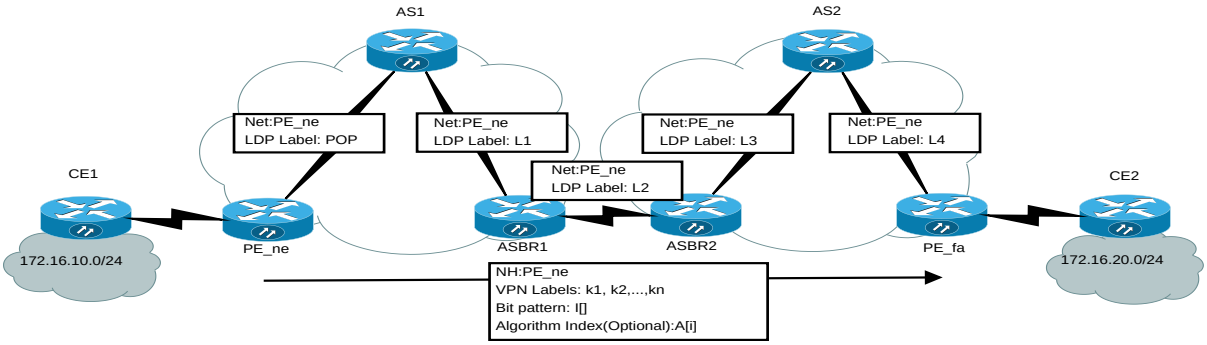


Fig. 1: Control-plane exchanges for model "C" [10]

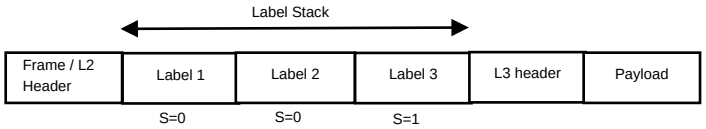


Fig. 2: Label stack using scheme outlined for model "C"

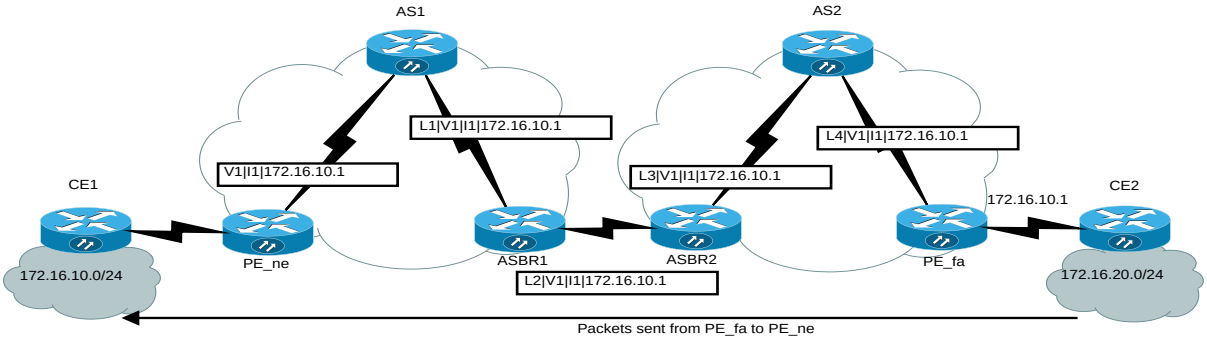


Fig. 3: Data-plane flow for model "C" [10]

Algorithm 4 Data-plane PE_{ne} algorithm

Require: None

```
1: Begin
2: packet = DP-ReceivePacket(Interface);
3: match = CheckFEC(packet);
4: if match == 0 then
5:    return; //no match
6: end if
7: label-in-packet=extractPacket(packet, LABEL);
8: inner-label=extractPacket(packet, INNER-LABEL);
9: hash-digest=calculateHash(A[i],packet);
10: first-label=hash-digest % |K|;
11: additional-label = process(hash-digest,I)
12: if label-in-packet ≠ first-label then
13:    error(); return;
14: end if
15: if inner-label ≠ additional-label then
16:    error(); return;
17: end if
18: DP-SendPacket(CE1, NULL, NULL, packet);
19: End
```

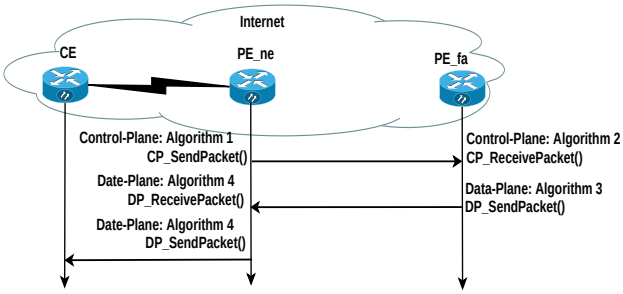


Fig. 4: A modified sequence diagram showing the applicability of the algorithms in the control and the data-plane.

The values in K need not be contiguous and can be randomly chosen from a pool of labels to remove coherence in the label space. Also the algorithms used could be either vendor dependent or a set of standard algorithms mapped the same way by the PE_{ne} and PE_{fa}. If the two PEs involved are from different vendors we assume that a set of standard algorithms are used. In order to avoid too many processing cycles in the line cards of PE_{ne} and PE_{fa}, the hash-digest is

calculated over a predefined size of the payload. The additional inner label is added to enhance protection against spoofing attacks. With an increased label size, an attacker spends more time to guess the VPN instance label for the site behind PE_{ne}. There could be two hash-digests that generate the same label. In this case, the two hash-digests are distinguished using the additional label. Collisions can be avoided by re-hashing or any other suitable techniques that are proposed in the literature [11]. If collisions exceed a certain number, then Algorithms 1 and 2 can be executed with a set of new labels.

A. Illustration

We now illustrate the label-hopping scheme. In Figure 1, using Algorithms 1 and 2, a set of labels are forwarded from PE_{ne} to PE_{fa}. The roles of PE_{ne} and PE_{fa} are interchanged for reverse traffic. Figure 2 shows a packet from the data-plane for model "C" with the proposed scheme. In Figure 2, "Label 1" refers to the outermost label, while "Label 2" refers to the label generated from the hash-digest and "Label 3" refers to the additional label that is generated as shown in Algorithm 3. This additional label, denoted by S, has a bottom of stack bit set; see Figure 2. These labels are stacked immediately onto the packet and the path labels for routing the packets to appropriate intermediary PEs are added. Figure 3 also shows these path labels used by the data packet to reach PE_{ne}.

Note that the labels that are exchanged need not be related with the services offered by the VPNs. A separate mapping can be maintained internally by the PEs. When the packet passes through the core of an intermediary AS involved in model "C", or through the network connecting the intermediary AS, the intruder or the attacker has the capability to inspect the labels and the payload. However, the proposed scheme prevents the attacker from guessing the right combination of the labels as the labels change with every data packet.

B. Computational time complexity

The computational time complexity of the algorithms executing at the control-plane is $O(1)$. The data-plane algorithms have a computational time complexity of $O(HashPacketSize)$. The packet size chosen for hashing could either be 64 or 128 bytes. Further control-plane exchanges are less frequent than the data-plane exchanges. In terms of processing, hashing small data sizes may not be an issue but frequently hashing every data packet increases the processing time. Hence, it would be of interest to reduce the computational time complexity of the data-plane to $O(1)$.

The most time consuming step in the data-plane algorithms is the hashing of data packets. We show how this hashing step can be removed by using the Timing over IP connection and Transfer of Clock (TicToc) [12] based Precision Time Protocol Label Switch Path (PTP-LSP) [13]. We discuss some important aspects of this algorithm.

IV. TicToc based label-hopping

If we use the TicToc based PTP LSP then a pre-calculated set of distinct values d_{ijk} for a specific time slot i, FEC j and a label index k could be exchanged over the control-plane periodically. These discrete values can then replace the hash values calculated in Algorithms 3 and 4 thereby improving speed-up. In this case, a few of the values from $d_{(i-1)jk}$ and $d_{(i+1)jk}$ must overlap with d_{ijk} forming a sliding window of distinct values. The sliding window is necessary to account for any latency in the clock information. In case $|d_{ijk}|$ is large then we can transfer a random seed for generating pseudo-random numbers R_{ij} which generates k values for every time instant i and FEC j. The algorithm for generating the pseudo-random numbers must, a-priori, be known to PE_{fa} and PE_{ne}. The sliding window of labels with the distinct values for three consecutive time slots is given in Figure 5.

The ports of the PEs must be configured to enable the functioning of the TicToc protocol. The rest of the configuration of the PEs is similar to the label-hopping schemed discussed in Section II-B.

As before, for each FEC or a set of FECs, a set of valid labels used for hopping in the initial time slot i is exchanged. These labels $D = (d_1; d_2; d_3; \therefore d_n)$ where $n \geq 1$ and, $k_i \neq k_j$ if $i \neq j$ are then configured in PE_{ne}. For the set of labels D time slices $TS = (TS_1; TS_2; TS_3 \therefore TS_n)$ are also exchanged. These time slices can be periodically changed and a new set of TS ranging from TS_1 to TS_n can be exchanged after a time duration of $TS\_Exchange\_Interval$ from time to time.

The complete sets of algorithms are given in the Appendix. The algorithm given for the control and the data-plane have a constant computational time complexity $O(1)$ while achieving the same objective of mitigating spoofing attacks. The main reason behind using TicToc is to synchronise the labels based on time. We could even consider the use of currently existing Network Time Protocol (NTP) [14] instead of TicToc to synchronise the labels. NTP is widely used to synchronise a device to Internet time servers or other sources. However, such a discussion is beyond the scope of this paper.

It should be noted that the algorithms protect against spoofing attacks. Replay attacks are still possible on systems implementing these schemes. In the next section, we show that Algorithms 1, 2, 3, and 4 can also be modified to mitigate replay attacks.

V. Mitigating replay attacks

In replay attacks, a valid data packet is replayed or delayed. Since the previous algorithm uses three consecutive time slots, an attacker can replay the packets within three time slots. In the hashing based algorithms the packet can be replayed many times until the labels are valid. Algorithms proposed in the previous sections cannot detect such attacks. Therefore to mitigate replay attacks we introduce a random seed. This random seed, henceforth referred as *Rseed*, generates pseudo random numbers which are used as the label for the time slots. We now discuss the modified algorithms in detail.

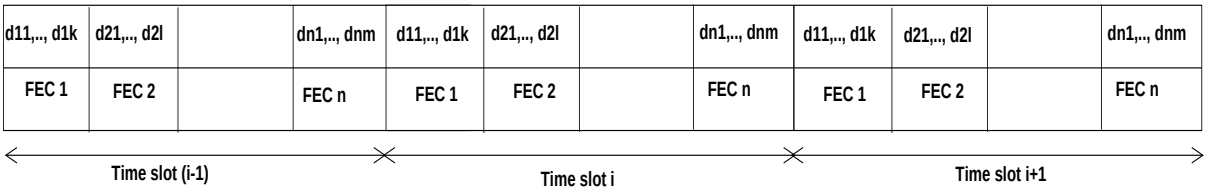


Fig. 5: Total distinct valid keys for a particular time slot i, various FECs j in a TicToc based protection against spoofing attacks. In case the keys are too large then a random seed which can generate the keys d_{ijk} can be exchanged. In this case, PE_{fa} and PE_{ne} must know the random number generation algorithm a-priori.

A. PE configuration

PEs that implement this scheme need extra configuration details in addition to those discussed in Section II-B. This includes the algorithm for pseudo random number generator, the random seed to be exchanged as well as the configuration of ports for implementing the TicToc protocol.

B. Control and Data-plane flow

As given in the TicToc based algorithms in the Appendix, the control-plane exchanges involve constructing a PTP LSP for deriving the clock at the PE_{ne} and PE_{fa} for the forwarding direction. Each pair of PE_{ne} and PE_{fa} knows the PTP port and corresponding PTP LSP used for the traffic. The PTP LSP is intended for providing the clocking between a pair of PE_{ne} and PE_{fa}. The clock or time-stamp derived from this PTP LSP is used in the data-plane to determine the valid label at that time instant. Upon validation, PE_{ne} uses the label information to forward the packets to the appropriate VPN service instance or site.

Once a data packet destined to the PE_{ne} arrives at the PE_{fa} the first-label is chosen using K, TS, and *Rseed*. A selected number of bytes from the payload is chosen as input to the hashing algorithm. The agreed bit-selection pattern is then applied on the hash-digest to obtain an additional label, which is then concatenated with the first label. Once PE_{ne} receives these packets it verifies both the labels. Note that in case hashing is not preferred we could use the predetermined set of labels as discussed in the Appendix. The details of the algorithm to mitigate spoofing as well as replay attacks are described below.

The PTP port number and the related PTP LSP information are assumed to be configured before any information exchange in the data-plane. Algorithm 5 forwards the FECs, their associated keys, a set of valid time slices, a random seed and the bit selection pattern for the inner labels from PE_{ne} to PE_{fa}. The packets are exchanged using Multi protocol External BGP (step 3).

Algorithm 6 runs at the PE_{fa}. It receives the packet and extracts information related with FECs, labels, time slices, random seed (steps $2-6$) and records them (steps $8-12$). It selects the hash algorithm based on the instance and uses the random seed value to generate pseudo-random numbers. It is assumed that both PE_{ne} and PE_{fa} will use the same pseudo-random number generator.

Algorithm 5 Control-plane PE_{ne} algorithm

Require: FEC[] Forward Equivalence Classes, K[] valid labels, TS[] valid time slices, A[i] hash algorithm instance, I[] the bit-selection pattern chosen for the inner label, Random seed *Rseed* which is used for generating the index into set K (set of labels), PTP port and PTP LSP information.

1: Begin
2: packet = makepacket(FEC,K, TS, A[i], I, Rseed);
3: CP-SendPacket(PE_{fa}, MP-eBGP, packet);
4: End

Algorithm 6 Control-plane PE_{fa} algorithm

Require: None

1: Begin
2: packet = CP-ReceivePacket(PE_{ne}); // from PE_{ne}
3: FEC[] = ExtractFEC(packet); // extract FECs
4: K[] = ExtractLabels(packet); // extract the labels
5: TS[] = ExtractTimeSlices(packet); // extract the time slices
6: Rseed = ExtractRandomSeed(packet); // extract the *Rseed* value.
7: selectHashAlgorithm(A[i]); // hash algorithm to use
8: RecordValues(FEC); // information for PE_{fa}
9: RecordValues(K);
10: RecordValues(TS);
11: RecordValues(I); // bit-selection pattern to be used
12: RecordValue(Rseed);
13: End

Algorithm 7 is implemented by PE_{fa}. Steps $2-6$ identify the current time slot. The keys for this time slot have already been exchanged in the control-plane. The algorithm works only if the label-hopping is enabled on the FECs. If the label hopping is enabled steps $13-26$ are executed. In step 13, the hash value of the packet is calculated. Steps $14-23$ manages the time slots. We assume that there are n time slots. If all the time slots are completed, we wrap around to time slot 0. A random number is generated and a key for the particular time slot is selected in step 24. The additional labels are created based on the previous identified bit pattern. The packet is then forwarded to the PE_{ne} (see steps $25-26$).

Algorithm 7 Data-plane PE_{fa} algorithm

Require: None

```
1: Begin // One time initialisation
2: CurrentTimeSliceIndex = 0;
3: CurrentMasterClock = PTP LSP Master Clock Timestamp;
4: CurrentTimeInstant = CurrentMasterClock;
5: NextTimeInstant = CurrentMasterClock
                     + TS[CurrentTimeSliceIndex];
6: End
7: Begin // repeated for every data packet
8: packet = DP-ReceivePacket(Interface);
9: match = CheckFEC(packet); // Is the algorithm enabled?
10: if match == 0 then
11:    return; // algorithm not enabled.
12: end if
13: hash-digest = calculateHash(A[i],packet);
14: if CurrentTimeInstant ≤ NextTimeInstant ((+ or -) configured seconds) then
15:    // do nothing;
16: else
17:    CurrentTimeSliceIndex++;
18:    if CurrentTimeSliceIndex == n then
19:       CurrentTimeSliceIndex = 0; // check to wrap around
20:    end if
21:    CurrentTimeInstant = NextTimeInstant;
22:    NextTimeInstant = CurrentTimeInstant
                         + TS[CurrentTimeSliceIndex];
23: end if
24: first-label = K[GenerateRandom(Rseed) % |K|];
25: addl-label = process(hash-digest,I)
26: DP-SendPacket(PE_ne, first-label, addl-label, packet);
27: End
```

Algorithm 8 has to take care of lead or lag in the clock. Since there could be a time-lag between sending and receiving packets, PE_{ne} has to maintain three random seeds. These include the random seed for the previous time slot and the current time slot. In case the time-slots have already wrapped once, the future random seed of the time slot is also stored. Steps $15-33$ takes care of this activity. The else part in steps $17-23$ stores the previous, the current and the next random seed (if it exists). The hashing should be applied on the packets and then the correct label must be chosen based on the random seed values. Steps $2-5$ does a one-time initialisation for the time slot. Functionality of steps $12-14$ and $35-42$ have been discussed in Algorithm 4.

The change in the algorithm to randomly pick up a label for the next time slot will help in avoiding man-in-the-middle attackers from synchronising with the time slots. The labels in the previous algorithms are predictable if a large number of packets were observed. The *Rseed* will generate values in lock step with the time slots at both the PE_{fa} and PE_{na}. This will

Algorithm 8 Data-plane PE_{ne} algorithm

Require: None

```
1: Begin // One time initialisation
2: CurrentTimeSliceIndex = 0;
3: CurrentMasterClock = PTP LSP Clock Timestamp;
4: CurrentTimeInstant = CurrentMasterClock;
5: NextTimeInstant = CurrentMasterClock
                     + TS[CurrentTimeSliceIndex];
6: Begin // For each packet
7: packet = DP-ReceivePacket(Interface);
8: match = CheckFEC(packet);
9: if match == 0 then
10:    return; //no match
11: end if
12: label-in-packet=extractPacket(packet, LABEL);
13: inner-label=extractPacket(packet, INNER-LABEL);
14: hash-digest=calculateHash(A[i],packet);
15: if CurrentTimeInstant ≤ NextTimeInstant ((+ or -) configured seconds) then
16:    // do nothing;
17: else
18:    CurrentTimeSliceIndex++;
19:    OldRseedIndex = RseedIndex;
20:    RseedIndex = (GenerateRandom(Rseed) % |K|);
21:    NextRseedIndex =
       LookAheadRseedIndex(GenerateRandom(Rseed)%|K|);
22:    RollbackRseed(Rseed by 1);
23:    if CurrentTimeSliceIndex == n then
24:       // check to wrap around
25:       CurrentTimeSliceIndex = 0;
26:    end if
27:    CurrentTimeInstant = NextTimeInstant;
28:    NextTimeInstant = CurrentTimeInstant
                         + TS[CurrentTimeSliceIndex];
29: end if
30: // Check if label used before in the previous, current
31: // or future time slot can be used
32: // Check with OldRseedIndex, RseedIndex
    // and NextRseedIndex
33: first-label = K[RseedIndex (+ or −1)];
34: additional-label = process(hash-digest,I)
35: if label-in-packet ≠ first-label then
36:    error(); return;
37: end if
38: if inner-label ≠ additional-label then
39:    error(); return;
40: end if
41: DP-SendPacket(CE1, NULL, NULL, packet);
42: End
```

prevent an attacker from synchronising with label changes and hence replay attacks could be avoided. The sequence diagram given in Figure 4 is valid for Algorithms 5-8.

Note that the changes to the label-hopping algorithms presented in this section can be applied to the algorithms given in the Appendix. This will ensure that the spoofing and replay attacks are mitigated close to wire speed. We do not discuss the details in this paper.

VI. Simulation

In this section, we present the simulation results on performance, comparing the various label-hopping technique including deep packet inspection where we encrypt and decrypt the complete packet.

Implementing the label-hopping algorithm for sets of FECs belonging to any or all VPN service instances may cause throughput degradation. This is because the hash-digest computation and derivation of the inner-label / additional inner label calculation can be intensive operations. We therefore compared our technique by choosing a part of the payload as input to our hashing algorithm.

We simulated our algorithm on a 2.5 GHz Intel dual processor quad core machine. We compared the performance of the label-hopping technique with a deep packet inspection technique where the complete packet was encrypted before transmission and decrypted on reception. The performance of the data-plane level algorithm on PE_{ne} is shown in Figure 6. Simulations without the use of TicToc schemes indicate that we were able to process 10 million packets per second when we used 64-byte for hashing on a payload of size 1024 bytes. For a hash using 128-byte, we were able to process approximately 6.3 million packets per second. However, with complete encryption and decryption of the packet, we were able to process only about 1 million packets per second.

The TicToc based algorithms given in the Appendix adds only constant time to the computation. There is an increase between $1-3\%$ in computation time depending on successful identification of label from the correct time slot. Therefore the results in Figure 6 show that the lines lie very closely to wire-speed performance.

When we combine label-hopping, TicToc and pseudo random number generation it adds approximately $2-5\%$ of additional computation time to the label-hopping algorithm. Since the difference in processing speeds is less than 5%, the performance with label-hopping lies very close to 64 and 128 bit hash lines in the figure. We were able to process approximately 9.6 million packets for a 64-bit hash and 6 million packets for 128-byte hash.

Based on the simulation results we suggest two solutions that can be implemented:

- The payload based label-hopping can be applied to specific VPN traffic between the PEs which are mission-critical and sensitive.
- The TicToc based label-hopping algorithm can be applied on those VPNs that have high link reliability and require line-rate operation.

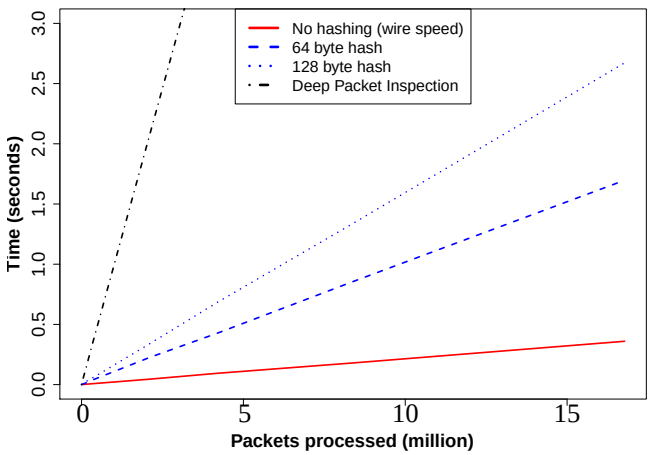


Fig. 6: Performance comparison of complete packet encryption and decryption with a 64, 128 byte hash on a payload of size 1024 bytes. The performance of the TicToc based label-hopping is very close to wire speed. Replay attack prevention adds approximately $2-5\%$ extra time to the algorithms.

In the next section, we consider some implementation issues that must be addressed while using the label-hopping methods.

VII. Implementation

In the PEs label-hopping can be enabled or disabled by using a look-up table. This look-up table can be used to efficiently implement the algorithms which can be programmed with an on or off bit to indicate whether the label-hopping scheme is deployed. In case the scheme is deployed, then the PEs compute inner labels 2 and 3 using Algorithm 8. If the packet is valid it is accepted, else it is discarded.

One concern of the scheme is to have a method to tackle the problem of fragmentation that can occur along the path from PE_{fa} to PE_{ne}. We can fragment the packet at PE_{fa} and ensure that the size of the packet is fixed before transmission. The other method is to employ the Path maximum transfer unit (Path-MTU) discovery process so that packets do not get split further into multiple fragments. If packets are fragmented this scheme fails. However, networks employ the Path-MTU discovery process to prevent fragmentation and hence this problem may not exist.

The proposed label-hopping method based on payload expects the same hashing algorithms to be used at both the PEs. If the vendors of the PE_{ne} and PE_{fa} are different then interoperability issues must be addressed. In our scheme, we chose the time instant that the packet leaves the PE_{fa} and this time instant serves as the variable component that the attacker cannot decipher. This requires the use of time synchronisation mechanism. This is provided by the PTP LSP.

VIII. Impact on applications

We briefly discuss the impact of the label-hopping method on a Deep Packet Inspection Engine (DPIE) and ISP billing. We show that by using label-hopping the workload on the DPIE can be reduced but the processing load on the ISP billing mechanism would increase.

A. Enhancing DPIE performance

Once a spoofed packet is detected at the PE_{ne} an error routine is called (see Algorithm 8). Such packets can be collected and periodically sent to a DPIE. Without the application of this algorithm, all the packets that are received by the customer site have to be sent to the DPIEs for possible attack detection. The proposed technique helps in filtering packets which would otherwise undergo deep packet inspection.

To mitigate attacks ranging from buffer overflow hack attempts to denial of service attacks such as tear drop attacks, the DPIEs record the packets in a secondary storage for inspection and correlation. The correlation engines also need computation power and memory, for deciphering and raising alarms to the about possible attack on specific VPN or a set of VPN sites. With reduced number of packets sent to DPIEs, the attack detection and correlation can be applied only to these filtered set of packets. It is important to note that if this label-hopping scheme was not adopted, some sort of DPIEs would have to be placed within the customer's network.

Another less preferred method is to have the DPIEs on the PE itself. With label-hopping scheme in place there is no need for having DPIEs on the PEs. The error packets can be spanned or replicated and sent to a suitable cluster of DPIE engines at the customer site for further correlation. An alternate solution for the ISP deploying the PEs could be to provide the first level of warning while the customer's hardware could do the rest of the mitigation and protection. This would be a co-operative solution between the customer and the ISP that reduces the time taken in the event of an attack. The label-hopping scheme would bring an extra level of protection.

B. ISP Billing

A concern of such a scheme is the billing related aspects for ISPs as the labels change periodically. Most of the ISPs use the labels to bill their customer. The modification to billing can be done as follows: the traffic statistics are collected for all the VPN labels as if they were separate labels. At the egress PEs, statistics are gathered for the data packet and labels coming towards it based on a set of labels that would be exchanged. This data can then be used along with the ASBR statistics (identifying the egress PE by the outer label) for billing purposes. The label-hopping scheme involves more processing for billing by the ISPs.

IX. Outlook

Today, there is reluctance among service providers to use MPLS VPN model "C" due to security concerns like spoofing and replay attacks. We propose methods to secure the Inter-Provider MPLS VPN model "C" data-plane by preventing spoofing, replay and other unidirectional attacks.

A. Contributions

In this paper, we proposed a label-hopping scheme for inter-provider BGP-based MPLS VPNs that employ MP e-BGP multi-hop control-plane exchanges. In such an environment without label-hopping, the data-plane is subject to spoofing and replay attacks. Spoofing attacks can be prevented by using the payload-based label-hopping scheme. A combination of label-hopping, TicToc and the use of pseudo-random numbers serve to mitigate replay attacks. The proposed schemes prevent the spoofed or the replayed packets from getting into a VPN site. Simulations show that the use of randomised labels for label-hopping along with TicToc can operate at line-rate. All the proposed schemes are computationally less intensive as compared to other encryption-based methods. One additional advantage, of the label-hopping scheme, is that the workload for deep packet inspection engines can be reduced. However, there would be an increase in computation time complexity for ISP billing. This trade off could be worth considering. We hope that the methods proposed will encourage ISPs to experiment with MPLS VPN model "C".

B. Avenues for future work

There are some cases where the label-hopping scheme cannot be used. For example, consider Equal Cost Multicast Path (ECMP) scenarios. In this case, a flow arriving at a router could choose any of the available equal cost paths to reach the destination. However, it is advisable that the flows of the same service choose a unique path out of the available equal cost paths. Otherwise, reordering of packets could occur at the receiving end as two equal cost paths may not have the same latency. For any real-time flows, reordering of packets introduces processing concerns at the receiver. The current practice to avoid reordering is to hash the flow labels so that flows of the same service are redirected through a specified unique path.

If flow hashing is done on Label Switching Routers (for pseudowires in RFC 6391), then the labels generated by the label-hopping technique might hash to different paths. Therefore reordering schemes have to be introduced at the receiver end. It is desirable to propose methods to solve this problem in the future.

Acknowledgements

The authors would like to acknowledge the UK EP-SRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC. The authors would like to thank Josh Rogers, James Uttaro, Robert Raszuk, Tal Mizrahi, Robert Raszuk, Greg Mirsky, Jakob Heitz and Bhargav Bhikkaji for the extensive and useful email discussions.

References

[1] S. Raman and G. Raina, *Mitigating Spoofing Attacks in MPLS-VPNs using Label-hopping*, Proceedings of the Eleventh International Conference on Networks (ICN 2012), pp. 241–245, ISBN: 978-1-61208-183-0.

[2] Y. Rekhter, B. Davie, E. Rosen, G. Swallow, D. Farinacci, and D. Katz, *Tag switching architecture overview*, Proceedings of the IEEE, vol. 85, no. 12, December 1997, pp. 1973–1983, doi:10.1109/5.650179.

[3] Advance MPLS VPN Security Tutorials [Online], Available: *http://etutorials.org/Networking/MPLS+VPN+security/Part+II+Advanced+MPLS+VPN+Security+Issues/*, [Accessed: 20th December 2012]

[4] M. H. Behringer and M. J. Morrow, *MPLS VPN security*, Cisco Press, June 2005, ISBN-10: 1587051834.

[5] S. Alouneh, A. En-Nouaary, and A. Agarwal, *MPLS security: an approach for unicast and multicast environments*, Annals of Telecommunications, Springer, vol. 64, no. 5, June 2009, pp. 391–400, doi:10.1007/s12243-009-0089-y.

[6] C. Semeria, "RFC 2547bis: BGP/MPLS VPN fundamentals", Juniper Networks white paper, March 2001.

[7] L. Fang, N. Bita, J. L. Le Roux, and J. Miles, *Interprovider IP-MPLS services: requirements, implementations, and challenges*, IEEE Communications Magazine, vol. 43, no. 6, June 2005, pp. 119–128, doi: 10.1109/MCOM.2005.1452840.

[8] C. Lin and W. Guowei, *Security research of VPN technology based on MPLS*, Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCSCT 10), August 2010, pp. 168–170, ISBN-13:9789525726107.

[9] B. Daugherty and C. Metz, *Multiprotocol Label Switching and IP, Part 1, MPLS VPNS over IP Tunnels*, IEEE Internet Computing, May–June 2005, pp. 68–72, doi: 10.1109/MIC.2005.61.

[10] Inter-provider MPLS VPN models [Online], Available: *http://mpls-configuration-on-cisco-ios-software.org.ua/1587051990/ch07lev1sec4.html*, [Accessed 20th December 2012].

[11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, 3rd edition, MIT Press, September 2009, ISBN-10:0262033844.

[12] Timing over IP and Transfer of Clock Work Group [Online], Available: *http://datatracker.ietf.org/wg/tictoc/*, [Accessed: 20th December 2012].

[13] Precision Time Protocol LSP [Online], Available: *http://www.nist.gov/el/isd/ieee/ieee1588.cfm*, [Accessed: 20th December 2012].

[14] D. L. Mills, *Computer Network Time Synchronization - the Network Time Protocol on Earth and in Space*, CRC Press, 2010. ISBN:1439814635,

APPENDIX

In this appendix, we show a constant time $O(1)$ algorithm for mitigating spoofing attacks in MPLS VPNs using time synchronized label-hopping. The time synchronization is achieved using the TicToc protocol. We assume that each (PE_{ne}, PE_{fa}) pair knows the PTP port and the corresponding PTP LSP used for sending and receiving time information. We now present the four algorithms which removes packet based hashing and uses time-based labels.

Algorithm 9 forwards the FECs, keys associated with the FECs and a set of valid time slices from PE_{ne} to PE_{fa}. The packets are exchanged using Multiprotocol External BGP in step 3. For every time slice instant this process is repeated. This is shown is steps $5-10$. The algorithm then wraps around and restarts from TS[0].

Algorithm 9 TicToc: Control-plane PE_{ne} algorithm

Require: FEC[] Forward Equivalence Classes, D[] valid labels, TS[] valid time slices, PTP port and PTP LSP information.

```
1: Begin
2: packet = makepacket(FEC, D, TS);
3: CP-SendPacket(PE_fa, MP-eBGP, packet);
4: Sleep(TS[1]);
5: For every time instant TS[i] starting from TS[2]...TS[n],
6: Begin
7: packet = makepacket(FEC, D);
8: CP-SendPacket(PE_fa, MP-eBGP, packet);
9: Sleep(TS[i]);
10: End
11: End
```

Algorithm 9 shows that PE_{ne} transfers the distinct values at every time instant TS[i]. PE_{ne} has to store the values for three consecutive time intervals to ensure that the latency involved in sending new labels to PE_{fa} is also accounted.

The values sent by the PE_{ne} are extracted from the packet in steps $3-5$. For example, the procedure ExtractLabelsAndAppend(packet) given in Algorithm 10 helps the PE_{fa} use the new labels received at every new time instant from PE_{ne}. All the values are recorded in steps $6-9$, for executing the verification activity when a packet arrives.

PE_{fa} initiates the synchronization activity in steps $1-6$ in Algorithm 10. Once a packet is received then the algorithm checks whether the FEC of the packet has label-hopping enabled in steps $9-12$. If enabled, the time instances are checked and a label is chosen for the current time index, added to the packet and sent; see steps $14-27$.

Extra work needs to be done at the level of the data-plane for managing the synchronization so that packets are not rejected. Hence, PE_{ne} can store values of D for three consecutive time slots. PE_{ne} synchronizes itself with the time slots of PE_{fa}. This is shown in steps $1-6$ and $14-25$ in Algorithm 12. If the label-hopping algorithm for the packet is enabled, then the

Algorithm 10 TicToc: Control-plane PE_{fa} algorithm

Require: None

```
1: Begin
2: packet = CP-ReceivePacket(PE_ne); // from PE_ne
3: FEC[] = ExtractFEC(packet); // extract FECs
4: D[] = ExtractLabelsAndAppend(packet); // labels
5: TS[] = ExtractTimeSlices(packet); // extract time slices
6: RecordValues(FEC); // information for PE_fa
7: RecordValues(D);
8: RecordValues(TS);
9: End
```

Algorithm 11 TicToc: Data-plane PE_{fa} algorithm

Require: None

```
1: Begin // One time initialization start
2: CurrentTimeSliceIndex = 0;
3: CurrentMasterClock = PTP LSP Master Clock Timestamp;
4: CurrentTimeInstant = CurrentMasterClock;
5: NextTimeInstant = CurrentMasterClock
                     + TS[CurrentTimeSliceIndex];
6: End // One time initialization end
7: Repeat forever
8: Begin
9: packet = DP-ReceivePacket(Interface);
10: match = CheckFEC(packet); // Is the algorithm enabled?
11: if match == 0 then
12:    return; // algorithm not enabled
13: end if
14: if CurrentTimeInstant ≤ NextTimeInstant (configured seconds) then
15:    // do nothing;
16: else
17:    // Move by next TS[i]
18:    CurrentTimeSliceIndex++;
19:    if CurrentTimeSliceIndex == n then
20:       // check to wrap around
21:       CurrentTimeSliceIndex = 0;
22:    end if
23:    CurrentTimeInstant = NextTimeInstant;
24:    NextTimeInstant = CurrentTimeInstant
                         + TS[CurrentTimeSliceIndex];
25: end if
26: label = Choose a label from CurrentTimeSliceIndex of D[];
27: DP-SendPacket(PE_ne, label, packet);
28: End
```

received label is recorded and searched in the array of labels that was already exchanged for that time slot. These activities are shown in steps $9-13$ and $26-30$. If the labels do not match then it is an error and hence the packet is discarded.

Algorithm 12 TicToc:Data-plane PE_{ne} algorithm

Require: None

```
1: Begin // One time initialization starts
2: CurrentTimeSliceIndex = 0;
3: CurrentMasterClock = PTP LSP Clock Timestamp;
4: CurrentTimeInstant = CurrentMasterClock;
5: NextTimeInstant = CurrentMasterClock
                     + TS[CurrentTimeSliceIndex];
6: End // One time initialization ends
7: Begin
8: packet = DP-ReceivePacket(Interface);
9: match = CheckFEC(packet);
10: if match == 0 then
11:    return; //no match
12: end if
13: label=extractPacket(packet, LABEL);
14: if CurrentTimeInstant ≤ NextTimeInstant (configured seconds) then
15:    // do nothing;
16: else
17:    // Move by next TS[i]
18:    CurrentTimeSliceIndex++;
19:    if CurrentTimeSliceIndex == n then
20:       // check to wrap around
21:       CurrentTimeSliceIndex = 0;
22:    end if
23:    CurrentTimeInstant = NextTimeInstant;
24:    NextTimeInstant = CurrentTimeInstant
                         + TS[CurrentTimeSliceIndex];
25: end if
26: // Note that the array D must be 3 times
    // larger in this case
27: first-label = Check whether the current label is in D[]
28: if label ≠ first-label then
29:    error(); return;
30: end if
31: DP-SendPacket(CE1, NULL, NULL, packet);
32: End
```

Some remarks about the algorithms are given below:

- The label size will include that of the additional label used in the label-hopping algorithms based on hashing.
- Due to non inclusion of additional label, bit selection pattern is not needed.
- A mechanism to handle packet losses may be used when the labels desynchronize.
- This algorithm can be implemented real-time and at nearly line-rate.

Wireless Automation Network Stability Evaluation

Radek Kuchta, Radovan Novotny, Jaroslav Kadlec, and Radimir Vrba
Faculty of Electrical Engineering and Communication,
Brno University of Technology
Brno, Czech Republic
kuchtar | novotnyr |kadlecja | vrbar@feec.vutbr.cz

Vladimir Sulc
MICRORISC, s.r.o.
Jicin, Czech Republic
sulc@microrisc.com

***Abstract* - This paper describes stability testing of new wireless communication platform for automation, summarizing currently used wireless modules and technologies from different vendors and comparing their main advantages and disadvantages. Based on the wireless modules study, we have chosen one of the promising technologies for deeper analysis in real home environment typical for wireless automation applications. Two typical test-cases for wireless communication parameter evaluation are described within the study in order to determine the limits of stability and low error rate. Further, the statistical method used is described and testing results along with the main advantages of tested wireless communication platform are discussed. This new wireless communication platform was designed and developed especially for home automation and telemetry projects and test case results prove the suitability of this wireless communication technology for home and office buildings environment.**

Keywords - Home Automation; IQRF; Wireless communication; Stability testing.

I. INTRODUCTION

With the advance of networking technologies and wireless communications, the popularity and the applications of Wireless Sensor Network (WSN) are increasing. Wireless connectivity has grown-up considerably in recent years and current trends show that the Wireless Sensor Networks will be an integral part of our lives, more than the present-day personal computers [1][2][3][4]. The driving forces for this development include the ease of access and flexibility in ad-hoc situation or temporary network setups supported by wireless connectivity, avoiding restrictive wired connection that relies on copper and fiber optic cabling.

Wireless sensor network consists of a quantity of spatially distributed wireless sensor nodes and actor nodes, which are densely deployed in wide areas. Wireless sensor networks use battery supplied sensing and I/O devices. A sensor node, also known as a sensor pod or a mote, is an autonomous subsystem in a wireless sensor network, which is capable to monitor physical or environmental condition or perform some data processing. The communication infrastructure intended to monitor and record conditions at diverse locations is an important subsystem of wireless sensor network. This subsystem sends that data to processors in the network. Each node in a sensor network is typically equipped with a radio transceiver. Sensor nodes gather information and send them to a central node (sink) or to actors for appropriate actions. Usage of Wireless Sensor Networks with low energy demands, low weight and intelligent networking features seems to be the most cost effective solution for many application areas. These devices incorporate wireless transceivers so that communication in short distances over a Radio Frequency (RF) channel is enabled. Wireless Sensor Networks can be used for many applications in various application fields such as automation of the buildings, machines, in the monitoring product quality or conditions at agriculture, medicine, and healthcare. Our goal related to this article was to find and validate technology for intelligent wireless network with low power consumption.

Section I and Section II are summary of existing technologies. A general overview of available wireless solution targeted to the small home automation applications and their main parameters and limitations is described in Section II. Following section defines case studies and issues of testing of the wireless communication platforms. Statistic tool and evaluation method is described in the Section IV followed by the measured results in Sections V and VI. Conclusion of final measured values and their short assessment is in the last Section VII.

II. STATE OF THE ART

Various wireless communication solutions are available from different vendors in the market. These solutions support different network topologies. Many of them are based on 802.15.4 [4] 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 (Table 1).

Probably, the most known standardized protocol that works on non-licensed bands is Zigbee [13]. It is a solution based on the IEEE 802.15.4 standard prepared by the Zigbee Alliance [6]. This standard was developed by consortium of industrial companies especially for building automation [7][8]. There are also special applications for industrial control, e.g., [9] [11] on remote access to the system and using small, independent wireless devices, [10][14][15] on building automation and telemetry applications, or an alarm system suitable for pervasive healthcare in rural areas [12]. Among the proprietary solutions, a reference can be made to the technology of MiWi launched by Microchip Technology Inc. [16]. 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. From other solutions available on the market, mention would be made, for example, of the solution promoted by Z-wave alliance [17][18].

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, lower reliability, and increased network complexity.

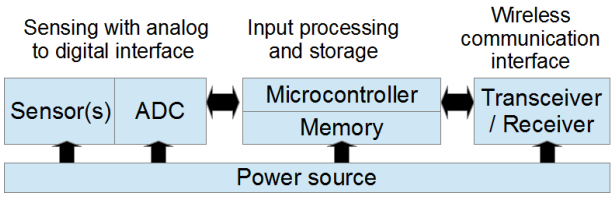


Figure 1. Block diagram of wireless sensor module

Wireless sensor module with RF measurement signal transmission is liable for collecting data from sensor when they could be recorded, configuration and switch on of other sensors and assuring other functions that are related to sensors. A sensor node is accomplished for gathering sensory information, performing some processing, and communicating with other linked nodes in the network. For example, the sensor module could be configured to allow a microcontroller to determine when a value of critical parameter has been reached or exceeded. Effectiveness of Wireless Sensor Networks (WSN) relies on the communication parameters of interconnected sensors' nodes, which are typically transmitting power, baud-rate, error-rate and their detection range or sensitivity to received signal.

These WSN technologies are determined especially for monitoring environmental and physical conditions, such as temperature, pressure, sound, vibration, humidity, and motion. WSNs applications are often used to perform many critical tasks and sensor networks applications have to meet strict rules and parameters to reliably and error-rate.

A failure of a component or components of a network may result in malfunction in the area of sensing, data processing, and communication. From this point of view it is necessary to evaluate the availability and reliability of application services as two important dependability factors [4].

A. *Mesh networking RF modules and improvement of reliability via redundancy*

RF modules are enabled for mesh networking and robust mesh networking topologies are preferred solution for developed applications. For example wireless sensor modules organized in a network in within the area of interest allow for monitoring of environmental conditions, events, or processes. A wireless mesh network is a communication network made up of nodes structured in a mesh topology. Multiple nodes cooperate to transport a message to its destination. Each of the nodes is able to cooperate with each other in transmitting packets through the network, especially in adverse conditions including the impact of powerful RF interferences.

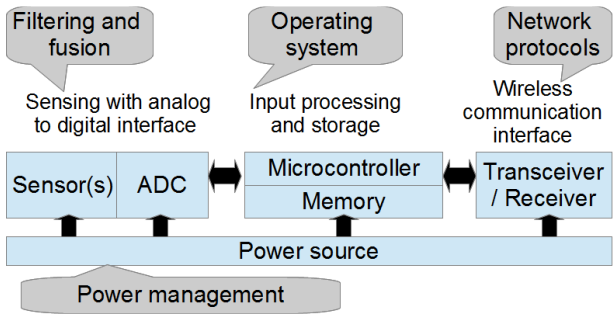


Figure 2. Important aspects of system architecture for wireless sensor networks

The mesh network features adaptability, self-configuring, flexibility, scalability and high-rate data transmission. This technology delivers scalable and self-healing properties: nodes are working in coordination with each other to create a network. Each node in the mesh network is connected to at least one other node and if a device or its link in a mesh network fails, data are sent around it via other redundant interconnections between network nodes. Nodes are talking to each other in a way that gets a message to a desired point using the applicable path and when a node is not in use, it will sleep using very little energy.

Table 1. Different RF modules of various manufacturers designed for operation in the license-free ISM bands

Producer, RF module type	Frequency range
Microrisc, TR-52B	868 MHz and 916 MHz
Texas Instruments, CC2500	2400-2483,5 MHz
Aurel Wireless, XTR-434	433,92 MHz
Sonmicro, SM130	13,56 MHz
Panasonic, PAN4555	2,4 GHz
Micrel, MICRF600	902-928 MHz
LS Research, ProFLEX01	2405 - 2480 MHz
Hope RF, RFM12-315	315, 433, 868, 915 MHz
Friendcom, FC-221/AP	433, 868 ,915 MHz
Shanghai Sunray Technology, SRWF-1028	403, 433, 470, 868, 915 MHz
Microchip Technology, MRF24J40MA	2,405-2,48 GHz
Aurel Wireless, RTX MID 3V/5V	433.92 MHz
Digi International Inc., XBee RF	2.4 GHz
Dorji Applied Technologies, DRF7020D13	433Mhz

The mesh topology offers multiple redundant communications paths throughout the network. This

redundancy, which is insensitive to the workload, enhances the overall reliability of the network: if one link for any reason fails, the network automatically routes messages through alternative paths. The degree of redundancy, which is a common strategy to enhance the reliability of systems, is essentially a function of node density. When one node can no longer work, the rest of the nodes can still communicate with each other, directly or through intermediate nodes. The redundancy improves the general reliability of the network simply by adding more node, links are more reliable without increasing transmitter power in individual nodes.

Table 2. Different RF modules of various manufacturers designed for operation in the license-free ISM bands

Producer, RF module type	Special firmware protocol	Modulation
Microrisc, TR-52B	IQMESH protocol	FSK
Texas Instruments, CC2500	No	2-FSK, GFSK, OOK, MSK
Aurel Wireless, XTR-434	No	FSK
Sonmicro, SM130	No	UART, I2C
Panasonic, PAN4555	SNAP® (Synapse Network Application Protocol)	Timer/Pulse Width Modulation (Tpm)
Micrel, MICRF600	No	FSK
LS Research, ProFLEX01	SIMPLE ProFLEX01	OQPSK
Hope RF, RFM12-315	No	FSK
Friendcom, FC-221/AP	Yes	3FSK
Shanghai Sunray Technology, SRWF-1028	No	GFSK, FSK
Microchip Technology, MRF24J40MA	No	FSK
Aurel Wireless, RTX MID 3V/5V	No	ASK
Digi International Inc., XBee RF	No	PWM0
Dorji Applied Technologies, DRF7020D13	No	FSK

B. Wireless networking protocols for security and efficiency

Network protocol is critical element of the networking and ensuring communication functions across implemented wireless network. Network protocols facilitate device identification and data transfer, and it is the special set of instructions that manages the communications among devices on a network. Network protocols play significant role because they are used for transmitting network packets across the network and are able to provide the altered types of paths to do the access to the network.

Network protocols also deal with the topologies of the network and can also take work in increasing the speed of data transmission. The aim is to use robust wireless communication protocols that are energy efficient and provide low latency. Principally a network protocol is defined as the set of regulations and rules of networking, which are needed for communicational process. The protocol software components are interfaced with a framework implemented on devices operating system. Different types of wireless networking protocols are utilized to achieve smoothly communication, and there are various competing schemes for routing packets across networks. As a base for unlicensed operation, standardized and proprietary protocols such as Bluetooth, ZigBee or WiFi are used and some RF modules manufacturers have their own protocols (see Table 2).

C. Transceivers operating in the ISM band

Governments regulate the use of various frequency bands and a license is necessary to operate in certain frequency bands. The industrial, scientific, and medical radio bands (ISM bands) are unlicensed frequency bands defined by the International Telecommunication Union Radiocommunication Sector (ITU-R). The ITU Radiocommunication Sector shows an important role in the worldwide organization of the radio-frequency spectrum. This part of the radio spectrum can be used without a license in most countries. License-free utilization is generally allowed in these bands though there are some differences in national regulations. These portions of the radio spectrum were originally reserved internationally for the use of radio frequency energy for industrial, scientific and medical purposes rather than communications. Despite the original purpose of ISM bands, there has been a large extension in its use in short-range, low-power, wireless communications platforms. Today the industrial, scientific and medical unlicensed Sub-GHz radio frequency bands are used for low-power wireless short-range wireless transfer of data at relatively low rates in point-to-point and more complex network topologies.

The employment of ISM equipment produces electromagnetic interference that disturbs radio communications in case of using similar frequency – the communications device must accept any interference produced by ISM equipment. Sharing the radio spectrum between various wireless devices that can work in the same environment may lead to oppressive interference interrupting radio communications or substantial performance degradation that makes use of the same frequency. The communication device working in these bands should reckon with consequences of the interference produced by ISM equipment because users do not have any regulatory protection. These undesirable effects are composed in densely occupied city areas with large numbers of wireless computer accessories, wireless remote controllers or

keyboards, cordless phones, Bluetooth devices, microwave ovens, and other devices which occupy the ISM band.

Table 3. Power consumption for various RF modules from different manufacturers

Producer, RF module type	Receiving current R_X	Transmitting current T_X	Sleep mode
Microrisc, TR-52B	35 µA - 13 mA	(14 – 24) mA	2 µA
Texas Instruments, CC2500	8,1 µA - 19,6 mA	(11,1-21,5) mA	400 nA
Aurel Wireless, XTR-434	(10-12) mA	(24-32) mA	100 nA
Sonmicro, SM130	180mA	180mA	30 µA
Panasonic, PAN4555	37 mA	30 mA	1 µA
Micrel, MICRF600	12 mA	10 / 23 mA	N/A
LS Research, ProFLEX01	25 - 35 mA	125 - 175 mA	8 µA
Hope RF, RFM12-315	10 mA	13 - 21 mA	0,3 µA
Friendcom, FC-221/AP	<65mA	<120mA	<25µA
Shanghai Sunray Technology, SRWF-1028	32-38mA	300-550mA	N/A
Microchip Technology, MRF24J40MA	19 mA	23 Ma	2 µA
Aurel Wireless, RTX MID 3V/5V	4,5-6,5 mA	13-20 mA	8 mA
Digi International Inc., XBee RF	50 mA	45 mA	< 10 µA
Dorji Applied Technologies, DRF7020D13	28 mA	35 mA	5 µA

D. Power consumption for transmit, receive and sleep modes

Building an energy efficient wireless network is one of the major efforts. Energy efficiency characteristics during routing are important in order to achieve energy savings. RF power consumption is very important parameter to be considered while designing battery powered systems. Typical module allows three different operating modes (Table 3). For the purpose of data receiving or transmitting there is a transmit mode and receive mode, and the sleep mode is designed in order to reduce the power consumption. Sleep mode can be activated by sending a sleep command, or by setting the module to return to sleep mode at every turn after data transmission.

In sleep mode, all of the system's interfaces shall support their necessary electrical levels and shall be able to wait for the next wake up period. When transmit mode is selected the module attempts to initialize an RF transmission and in combination with receive mode initiates an RF connection with other modules.

III. CASE STUDY DESCRIPTION AND PROBLEM DEFINITION

Small battery operated wireless sensor nodes are in our network used for automatic monitoring in the system. This application not only expects wireless signal coverage but also needs undisrupted service and reliable connectivity. The key aspect of wireless channel is the monitoring and evaluation of the channel quality. Most of the models of radio wave propagation involve questions related to the "free space" radio wave propagation [19].

Radio waves emit from a point source of radio energy, traveling in all directions. Obstacles such as physical and structural components of a building, furniture and fixed or movable structures, or the ground can impact signal propagation paths. Especially ferrous materials, such as steel and iron, can drastically alter signal propagation characteristics, communication distances, link quality, and many other factors [20].

Reflection, diffraction and scattering cause radio signal distortions and give rise to signal fades, as well as additional signal propagation losses. Indoor use of wireless systems creates the necessity for evaluation of indoor radio (RF) propagation. Any obstacles in the pathway would be harmful to RF transmission; radio signals penetrate obstacles in ways that appear hardly predictable. The final composite signal is made up of a number of components from the various sources of scattered and diffracted signal components or reflections from different directions.

To better understand this effect in our case study we at first evaluated the communication characteristics when sensor nodes were placed in various locations and distances. Absorption of RF energy resulted in loss of signal strength and reduced transmission distances. RF signals from wireless sensor nodes are air radiating from a transmitter and propagating through a medium in all directions. We need to understand the communication distance of individual nodes as well as to evaluate how and where to install the nodes.

The WSN in this case study is based on the IQRF (Intelligent Radio Frequency) wireless communication platform for industrial and home automation. This is the technology that was specifically developed for wireless sensor mesh networks by Microrisc company [21]. Typical application scenario of home automation with IQRF communication technology for a smart house is shown in the Fig. 3. Wireless sensor network as part of network infrastructure is depicted in Figure 3, sensing and controlling is accomplished by means of connected sensors and actuators. Upon detecting a relevant stimulus (signal) by sensors in the sensor field, sensors send reports to a sink node.

Networking collectively many of cheap wireless sensor nodes lets accurately evaluate a remote processes by utilization of the data from the individual nodes [25].

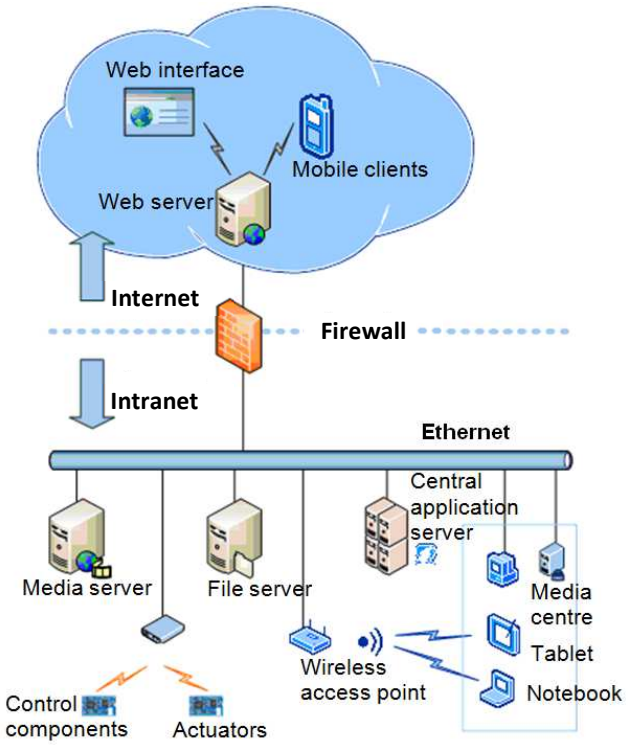


Figure 3. Block diagram of the telemetry and control for smart house

The main parts of the platform are covered by Czech and US patents [22][23][24]. For our experimental purposes, the standard IQRF components and development tools have been used. This wireless solution could be used for wireless connectivity necessary for telemetry, remote control, displaying of remotely acquired data, connection of more equipment and building automation. The implementation of IQRF transceiver modules works in non-licensed communication bands, license-free ISM bands 868 MHz in EU, 916 MHz in US, 433 MHz in EU, US and other countries.

Basic features of the IQRF communication platform are especially extra low power consumption (1 µA in the sleep mode and 35 µA in the on-line mode), available networking functions, programmable RF power up to 3.5 mW, SW selectable in steps, up to 170 m communication range, 15 kb/s (optionally 100 kb/s) RF bit rate. A transceiver module is the basic communication component needed for realization of wireless RF connectivity and can work as a node or a network coordinator. The IRQF modules could be integrated into any electronic device via SIM (Subscriber Identity Module) card connector. The low power consumption predetermines these modules for battery powered applications. The transceiver module is equipped with the IQRF operating system supporting functionality for the user application. There are RF functions for transmitting, receiving, network bonding, routing, main parameters configuration, EEPROM access functions, and IIC (Inter-Integrated Circuit bus) and SPI (SPI - Serial Peripheral Interface) communication functions. Data processing, for example, encoding, encryption, checksums, adding headers, is evaluated automatically by IQRF operating system during the communication. The other functions of operating system are three buffers and some other auxiliary functions. IQRF operating system is buffer-oriented and allows sending up to 32 bytes in one packet.

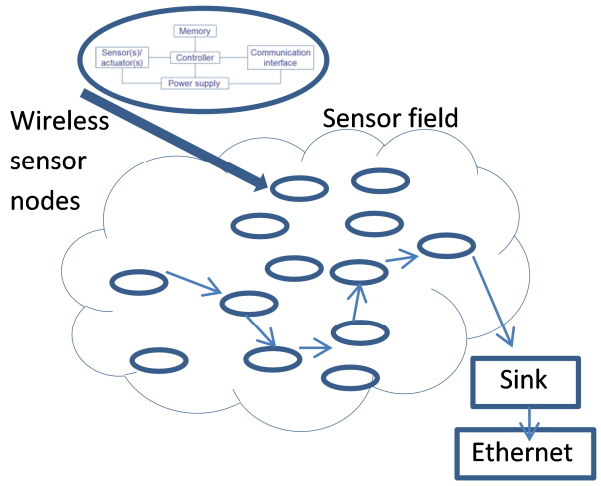


Figure 4. Wireless sensor network as part of network infrastructure

This application case study demonstrates simple data collection from wirelessly connected sensors. The network used for our experiment consists of one coordinator and a set of communication units. This is the basic star network topology where a sensor network is created around a core coordinator. The packet is wirelessly sent by operating system to the coordinator and the quality of the communication is statistically evaluated.

IV. CASE STUDIES AND USED DATA EVALUATION METHOD

For wireless communication parameters measurement were used two basic measurement criteria. The first one examined communications' parameters and wireless technology limits under typical building environment with the set of rooms separated by the plasterboard walls and the second set of measures was done in a long hall without any physical obstacles to test free space signal propagation.

A. Statistical description of the wireless network stability

The binomial distribution $B(n,p)$ with parameters n and p gives the discrete probability distribution of independent observations by the number of observations in the group that represent one of two outcomes. This distribution describes the behavior of a count variable X if the number of samples n is fixed, each sample represents "success" or "failure", each sample is independent, and the probability of "success" p is the same for each outcome.

The binomial distribution gives the approach to dependability evaluation for wireless communication. We expect that in the stable wireless network each type of outcome has a fixed probability, and by evaluation of the proportion of individuals in a random sample we could evaluate the stability of the network. They are the sequences of independent transfers in the communication model with two possible outcomes ("success" or "failure").

B. Statistical description of the wireless network stability

We extract samples of a certain size from the ongoing Wireless Sensor Network in our case study related to the stability testing of the channel quality. These are the sequences of independent transfers with two possible outcomes ("success" or "failure") in this experimental situation. The fraction or proportion of "failure" items can be expressed as a decimal or as a percent (when multiplication by 100 is used).

From the statistical point of view, the number of failures is the random variable. Common-causes and special-causes are the two distinct origins of variation. One common-cause variation is the noise within the system and is inherent to the process. It could be removed by making modifications to the process. Special-causes are unusual, not previously observed variation, which is inherently unpredictable. There are only common-causes in the stable system and the statistical monitoring and control could be used for stability evaluation.

Each run that is accomplished is then a realization of a Bernoulli random variable with parameter *p*. The binomial distribution *B*(*n*,*p*) with parameters *n* and *p* gives us the discrete probability distribution of these independent observations. If a random sample of *n* units of transfer realization is selected and if *k* is the number of units that are nonconforming, the *k* follows a binomial distribution with parameters *n* and *p* according to following equation

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall\, 0 \le k \le n \qquad (1)$$

We expect that in the stable wireless network each type of outcome has a fixed probability and by determining of the proportion of individuals in a random sample we can evaluate the stability of the network in setting condition. The np-chart such as Shewhart control chart with underlying binomial distribution can be used for the stability evaluation [26]. The sample size is constant and the amount of the unsatisfactory is plotted in a graph along with regulation limits. The regulation limits are defined as

$$n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})}, \qquad (2)$$

where *n* is the sample size and $\bar{p}$ is the estimation of the long-term mean. Rational subgroups for our testing are composed of the transfer of packets under essentially the same experimental conditions.

V. STABILITY EVALUATION OF INDOOR RF PROPAGATION (THE CASE STUDY OF MORE RF PROPAGATION OBSTACLES)

In this application case study, there were five transmission units in five various rooms each separated by the plasterboard partitions. There are two changed factors in this experiment: eight various levels of transmitting power and the time of the day. In each run, 80 data transfers were executed each of which consisted of 500 data frames. The selected scenarios cover most typical usages of this technology.

The number of failures in the communication was then evaluated. Failure in our case study indicates that the data frame was not received or there was a bit error in the data transfer (mismatched CRC). The results from this first experimental case study are summarized in Figure 6. There are two factors influencing the results. The mark (a) in the graph highlights the independence of the number of failures on the RF power. For a distance that is higher than 5 meters, it is necessary to optimize the RF power value. The mark (b) in the graph highlights the special-causes variation. The influence of parasitic effects such as the interference from microwave, treadmill, vacuum cleaner etc. has not been tested. These parasitic sources of wireless signal interference have unpredictable and hardly tested behavior which rapidly varies in every application conditions.

Experimental results were evaluated by using the np control charts (see Fig. 3 and Fig. 5). An np-chart is a plot of the number of failed items observed in a sample where *n* is the sample size and *p* is the probability of observing a defective item when the system is in control without affection of special cause variation. The statistical distribution of the number of failed items is assumed to be binomial. The observed number nonconforming (NP) is plotted against the control limits (UCL – Upper Control Limit, LCL – Lower Control Limit), which are statistically determined. These limits are usually calculated as three standard deviations from the mean, so there is around a 99.73 percent probability that a data point representing actual value of tracked parameter will be within those limits in the case of stability.

For the purpose of statistical evaluation of this wireless communication, experiments were used np control charts. The results of this analysis are summarized at the control charts (Fig. 6 and Fig. 8). If a data distribution is approximately normal the fluctuation of the points between the control limits (UCL, LCL) is due to the common cause variation. Then about 99.7 percent of the data values are within three standard deviations: $\mu \pm 3\sigma$, where μ is the arithmetic mean and σ is the standard deviation. Any points outside the control limits related to the six standard deviation empirical rule could be attributed to a special-cause variation. There are some cases where special-causes are affecting the results. Out of control points are marked as “1”. Overall interpretation of created np control charts for this part of experiment leads to these conclusions:

- The higher distance between the transmitter and receiver is in the relation to the special-cases variation existence and communication failure.
- The higher RF power gives the higher probability for wireless communication without failures.

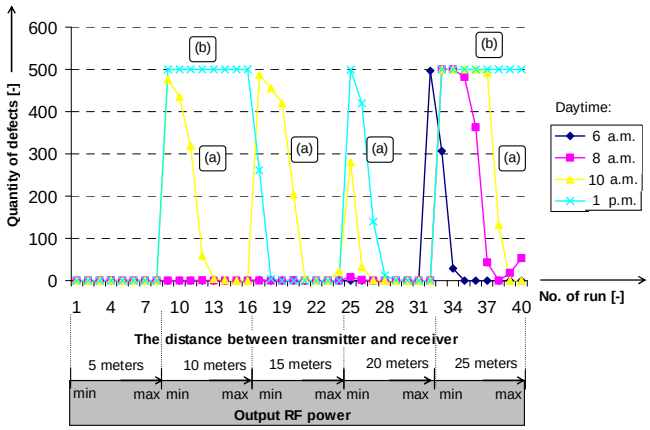


Figure 5. Wireless stability evaluation (the case study of more RF propagation obstacles)

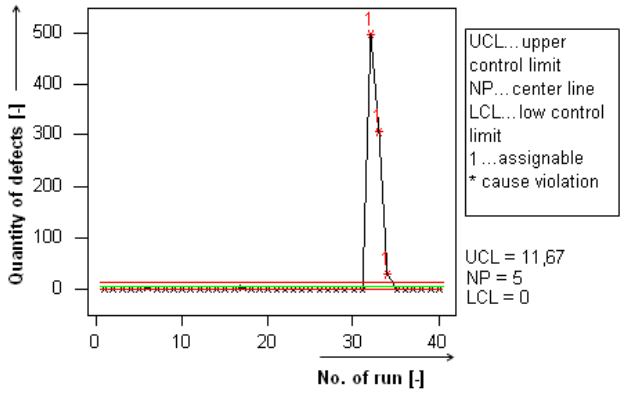


Figure 6. NP chart of wireless stability evaluation for the measurement at 6 p.m. (the case study of more RF obstacles)

VI. STABILITY EVALUATION OF INDOOR RF PROPAGATION (THE CASE OF THE FREE SPACE RF PROPAGATION)

In this part of the experiment, there were five transmission units in five various places in a region, which is free of all objects that might absorb or reflect radio energy. Eight various levels of the RF power and the various daytime are changed in this experiment. In each run, there was 80 data transfer execution, which each consists from 500 data frame. The number of failures in the communication system in this configuration was then evaluated. The results from this part of experimental case study are summarized in the Fig. 7.

We could see that there are some communications problems related to the setting of the RF power. The RF power needs the optimization according the distance between the transmitter and receiver. These situations are in the Fig. 7 depicted by mark (a).

Communication with the fifth transmitter unit located in the distance 25 meters for the receiver is affected by special cause variation in this case. This is the limiting distance that the signal is able to penetrate at the building environment in this experiment configuration. The reason for this is a hardware solution based on the used inbuilt antenna. Transmitting power is independent on the voltage level from the battery. If critically low value of battery voltage is reached transmitter stops operation and goes to switch-off state. Also temperature will not affect the result, because of the automated periodical crystal recalibration. Other parts of the transmitted do not have high temperature dependence.

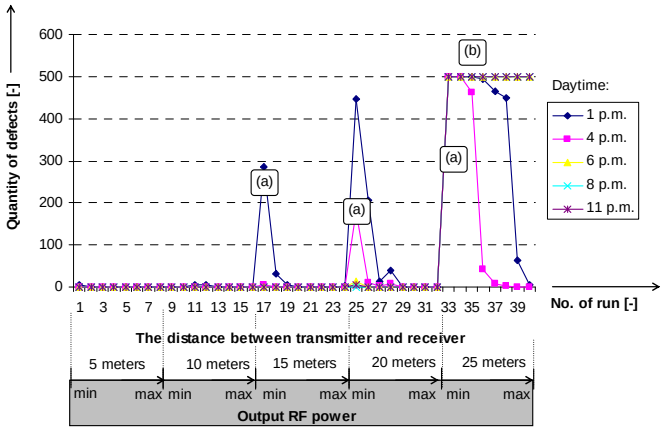


Figure 7. Wireless stability evaluation (the case study of the free space propagation)

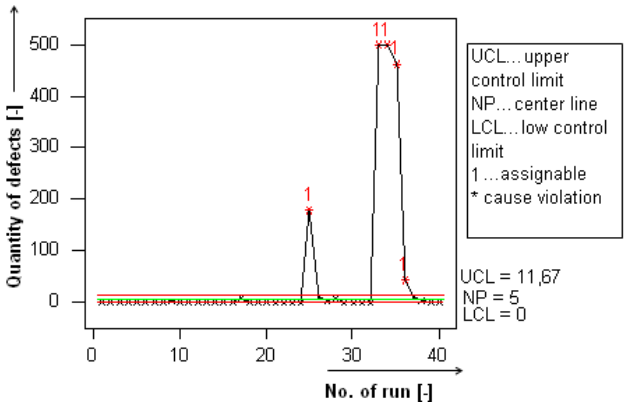


Figure 8. NP chart for wireless stability evaluation (the case of the free space propagation, 4 p.m.)

The mark (b) in the Fig. 7 is related to the communication failure and the special cause variation case.

The number of communication problems in comparison to the case of more RF propagation obstacles is smaller. The requirements for higher RF power are smaller and the overall stability is better. The higher RF power gives the higher probability of wireless communication without failures.

VII. CONCLUSION AND FUTURE WORK

Our goal related to this article was to find and validate technology for intelligent wireless network with low power consumption. 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. As a typical representative of the low-cost wireless communication technology IQRF presents ideal solution for home automation and office or light industry applications. As such, this platform was designed especially for home automation and telemetry applications.

This paper describes stability testing of the IQRF wireless platform in some real cases. We have focused on the long-term stability test and the verification of the parameters required for implementing the network. Proposed case studies have proved suitability of this technology to typical application scenarios, test their real communication parameters under buildings environment and determine limits. We are using real test cases i.e., sequence of events (actions) whose purpose is to find defects in a communication transfer. Based on the statistical result of measured data analysis, optimal node distance and output RF power to communication defects ratio can be set. Output RF power influences power consumption and then operation time. Optimal combination of distance and output RF power in specific operation conditions under different environments is therefore highly needed and can significantly improve operation time and minimize communication failures.

Real tests proved wireless communication abilities of IQRF, which fits to the requirements for usage in home automation and telemetry applications and also in the currently developed automatic stochastic system.

Our future work covers implementation and testing of selected wireless technology under real application conditions in the wide wireless sensors network.

ACKNOWLEDGMENT

This research has been supported by ARTEMIS JU in Project No. 100205 POLLUX - Process Oriented Electronic Control Units for Electric Vehicles Developed on a multi-system real-time embedded platform, by the Czech Ministry of Industry and Trade in projects FR-TI3/254 OPT - Open Platform for Telemetry and by the CZ.1.05/1.1.00/02.0068, OP RDI CEITEC - Central European Institute of Technology.

REFERENCES

[1] Kuchta, R.; Novotný, R.; Kadlec, J.; Vrba, R.; Sulc, V. Wireless Home Automation Network Stability Testing. In The Eleventh International Conference on Networks ICN 2012. Saint Gilles, Reunion Island: IARIA, 2012. s. 39-43. ISBN: 978-1-61208-183- 0.

[2] Akyildiz, F., Weilian, S., Sankarasubramaniam, Y., and Cayirci, E. A Survey on Sensor Networks, IEEE Communications Magazine, pp. 102-114, August 2002.

[3] Akyildiz, F., Wang, X., and Wang, W. Wireless mesh networks: a survey, Computer Networks, no. 47, pp. 445-487, January 2005, doi:10.1016/j.comnet.2004.12.001.

[4] Taherkordi, A., Taleghan, A., and Sharifi, M. Dependability Considerations in Wireless Sensor Networks Applications, Journal of Networks, vol. 1, no. 6, pp. 28-35, November 2006.

[5] Naris, L. and Benedetto, G. 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.

[6] ZigBee: (2009, May) ZigBee Aliance Web Pages. [Online]. HYPERLINK "http://www.zigbee.org" http://www.zigbee.org , Last accessed on 27 December 2011, <retrieved: 1, 2011>.

[7] Evans-Pughe, C. Bzzzz zzz [ZigBee wireless standard], IEE Review, pp. 28-31, March 2003, 0953-5683.

[8] Gill, K., Yang, H., Yao, F., and Lu, X. A ZigBee-Based Home Automation System, IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, pp. 422-430, May 2009.

[9] Gill, K., Yang, H., Yao, F., and Lu, X. A zigbee-based home automation system, Consumer Electronics, IEEE Transactions on, pp. 422-430, March 2009, 0098-3063.

[10] Edgan, D. The emergence of ZigBee in building automation and industrial control, Computing & Control Engineering Journal, pp. 14-19, April-May 2005.

[11] Zualkernan, A., Al-Ali, R., Jabbar, A., Zabalawi, I., and Wasfy, A. InfoPods: Zigbee-based remote information monitoring devices for smart-homes, Consumer Electronics, IEEE Transactions on, pp. 1221-1226, August 2009.

[12] Casas, R., Marco, A., Plaza, I., Garrido, Y., and Falco, J. ZigBee-based alarm system for pervasive healthcare in rural areas, Communications, IET , pp. 208-214, February 2008.

[13] Poole, I.: What exactly is... ZigBee?, Communications Engineer , pp. 44-45, August-September 2004.

[14] Ciardiello, T. Wireless communications for industrial control and monitoring, Computing & Control Engineering Journal , pp. 12-13, April-May 2005.

[15] Gomez, C. and Paradells, J. Wireless Home Automation Networks: A Survey of Architectures and Technologies, IEEE COMMUNICATIONS MAGAZINE, vol. 48, no. 6, pp. 92-101, June 2010.

[16] Flowers, D. and Yang, Y. MiWi Wireless Networking Protocol Stack, 2010. [Online]. HYPERLINK "http://www.newark.com/pdfs/techarticles/microchip/AN1066_MiWi AppNote.pdf", Last accessed on 27 December 2011, <retrieved: 1, 2011>.

[17] Gomez, C. and Paradells, J. Wireless home automation networks: A survey of architectures and technologies, Communications Magazine, IEEE , pp. 92-101, June 2010.

[18] Walko, J. Home Control, Computing & Control Engineering Journal, pp. 16-19, October-November 2009.

[19] Rappaport, T. Wireless Communications: Principles and Practice. Prentice-Hall, Englewood Cliffs, NJ: IEEE Press (The Institute of Electrical And Electronics Engineers, Inc.), 1996, ISBN: 0-7803-1167-1.

[20] Sun, Z. and Akyildiz, F. Channel Modeling and Analysis for Wireless Networks in Underground Mines and Road Tunnels, IEEE Transactions on Communications, vol. 58, no. 6, pp. 1758-1768, June 2010, ISSN: 0090-6778.

[21] Microrisc (2009, May) Microrisc Web Page. [Online]. HYPERLINK "http://www.microrisc.cz/new/weben/index.php" http://www.microrisc.cz/new/weben/index.php, Last accessed on 27 December 2011, <retrieved: 1, 2011>.

[22] Šulc, V. 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.

[23] Šulc, V. 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.

[24] Šulc, V., Kuchta, R., Vrba, R. IQMESH implementation in IQRF wireless communication platform, In 2009 Second International Conference on Advances in Mesh Networks, pp. 62-65, 2009, ISBN 978-0-7695-3667-5.

[25] Heintelman, W., Chandrakasan, A., & Balakrishan, H. An Application-Specific Protocol Architecture forWireless Microsensor Networks. IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 1, NO. 4, pp. 660-670, October 2002 .

[26] Ho, L., Costa, A. Monitoring a wandering mean with an np chart. Producao, v. 21, n. 2, p 254-258, June 2011.

NGS workflow Optimization using a Hybrid Cloud Infrastructure

Lorenzo Mossucca, Olivier Terzo, Klodiana Goga
Infrastructure and Systems for Advanced Computing (IS4AC)
Istituto Superiore Mario Boella (ISMB)
Torino, Italy
mossucca@ismb.it, terzo@ismb.it, goga@ismb.it

Andrea Acquaviva, Francesco Abate, Rosalba Provenzano
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy
andrea.acquaviva@polito.it, francesco.abate@polito.it

***Abstract*—E-science applications involve great deal of data, to satisfy these processing requests, distributed computing paradigms, such as cluster, Grid, Virtual Grid, Cloud Computing, and Hybrid System are growing exponentially. Existing computing infrastructures, software system design, and use cases have to take into account the enormity in volume of requests, size of data and computing load. In Bioinformatics field, such as in Next Generation Sequencing technology, in order to have more accurate analysis, it increases the amount of data to process. A new protocol for sequencing the messenger RNA in a cell, known as RNA-Seq, produces millions of short sequence fragments in a single run. These fragments can be used to measure levels of gene expression and to identify novel splice variants of genes. The proposed solution allows to make the system scalable and flexible reducing elaboration time. The first aspect covers reverse engineering of a fast splice junction mapper for RNA-Seq reads called TopHat in order to make parallelizable tasks and the second aspect concerns the development of hybrid architecture integrating a Grid and a Virtual Grid Environment.**

***Keywords*-grid computing; cloud computing; virtual; next generation sequencing; hybrid architecture.**

I. INTRODUCTION

Next Generation Sequencing (NGS) technologies, also known as second generation, have revolutionized biology and genomic research with the ability to draw from a single experiment a larger amount of data sequence [1] than with the previous technology known as Sanger Sequencing. DNA sequencing includes several methods and technologies that are used to determine the order of the nucleotide bases (adenine, guanine, cytosine, and thymine)in a molecule of DNA. The main novelty introduced by the NGS platform is to obtain smaller fragments from the molecules of DNA/RNA, called read, which are sequenced in parallel thus reducing the processing time [2], [3]. Aberrant mutations in the RNA transcription, as chimeric transcripts, are on the base of various forms of disease and NGS proved to be extremely helpful in making the detection of these events more accurate and reliable. However, even if from the biological point of view NGS technology leads to new exciting perspectives spreading an incredible amount of data, it raised new challenges in the development of tools and computing infrastructures. A NGS machine generates millions of reads in a single run that must be successively elaborated and analyzed. TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. It is built on the ultrafast short read mapping program Bowtie. TopHat finds splice junctions without a reference annotation. By first mapping RNA-Seq reads to the genome, TopHat identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. Using this initial mapping, TopHat builds a database of possible splice junctions, and then maps the reads against this junction to confirm them. Out objective is to offer to biologist private infrastructures to conduct their research and to respond to the ever evolving needs of NGS users. Cloud Computing is rapidly emerging as an alternative platform for the computational and data needs of our community. Biologists are already using the Amazon Elastic Cloud Computing (EC2) infrastructure for their research. In some situations where such a sensitive data are involved, for privacy and data security issue, it is preferable to use a number of instances of a tailored Virtual Machine (VM) than submitting jobs to the own existing infrastructure. Grids appear mainly in high performance computing environments. In this context, several of off-the-shelf nodes can be linked together and work in parallel to solve problems, that, previously, could be addressed sequentially or by using supercomputers. Grid Computing is a technique developed to elaborate enormous amounts of data and enables large-scale resource sharing to solve problem by exploiting distributed scenarios. The main advantage of Grid is due to parallel computing, indeed if a problem can be split in smaller tasks, that can be executed independently, its solution calculation speed up considerably. The paper is organized as follows: In Section II, main issues of the NGS technology are discussed that led to the realization of this project. Section III explains biological background, software and tools used. Section IV, designed architecture is shown: grid and virtual environment and schedulers functionalities. Section V is related to the test performance. The last section draws the conclusion and directions for future works.

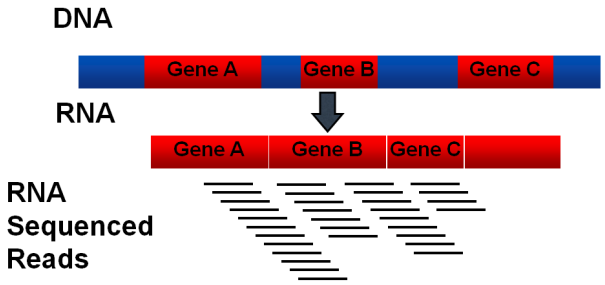


Figure 1. Alignment Phase.

II. Motivation

The amount of data produced with NGS technology is a positive factor that on a hand contributes to make studies more accurate and reliable identification of mutations in aberrant splicing events, fused genes, on the other hand, open new challenges in the development tools and infrastructure that are able to do the post-processing of data produced in a powerful and timely fashion [4]. A NGS data sample consists of millions of reads, and in a classic situation, with only one workstation available, the time needed to obtain the output increases significantly. In such a context, this computing infrastructure allows to improve overall system performance optimizing the use of resources and increasing the system scalability. The alignment is a process in which each mapping reference is made to read independently from the other reads, and this means that you can perform a parallel analysis of data. Even if the alignment is a very basic operation, due to the great number of data involved in the process, the computational effort in this phase is very high. This scenario recalls for the need of developing computing infrastructures presenting high performances CPU capability and memory availability [5].

III. Next Generations Sequencing

The aim of splicing tools is to track inside of the analyzed samples, the junctions between exons generated as a result of the process of splicing. This process allows that from the pre-mRNA introns are deleted and the remaining exons are joined together to form the mature RNA. The junction between exons spliced RNA is called splicing point, and thanks to the identification of these points it allows to figure out if events have occurred or whether alternative splicing events have occurred or whether we are in the presence of fused genes. There are several tools for splicing such as Supersplat, Splitseek and TopHat, in this section a brief overview for each tool is shown. Supersplat is a method for unbiased splice junction discovery through empirical RNA-seq data. Using a genomic reference and RNA-seq high throughput sequencing dataset, it empirically identifies potential splice junctions at a rate of approximately 114 million read per hour. The method depends upon the availability of previously known splice junctions on which to train the algorithm, and, when finding novel potential splice junctions, is biased toward those which are similar to its training data. In scoring novel potential splice junctions, the algorithm is biased toward canonical terminal dinucleotides, scoring those which conform to these biases higher than ones that do not. While, in general, these biases may prove to be correct, many potential splice junctions which do not conform to these rules threaten to remain unidentified. SplitSeek performs a number of analysis steps to predict the exon boundaries. First, all instances of split reads are found, and their genomic positions and nucleotide sequence are recorded. They comprise the initial set of candidates, and all resulting splice events will be found among these. However, many reads exist in which the junction is located in one of the anchors rather than in the gap. To identify such additional junction reads, it allows to scan all reads in which only one of the anchors was aligned. If such an anchor can be extended to the exact position as a previously identified candidate junction, and the sequence in the two reads aligns perfectly within the first five bases of the other exon, then the read is considered to confirm the junction. SplitSeek can find junction reads in which as few as five bases overlap with the other exon. In the final step, all identified junction reads are grouped, and user-defined cut-offs are applied to obtain a final set of exon boundaries. Because this method is unbiased, it will report all types of events in which a read must be split to match the reference genome, including small insertions and deletions.

A. TopHat algorithm

TopHat [6] is a fast splice junction mapper for RNA-Seq reads that aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. TopHat is a collaborative effort between the University of Maryland Center for Bioinformatics and Computational Biology and the University of California, Berkeley Departments of Mathematics and Molecular and Cell Biology. On the market there are several NGS platforms, Roche 454, Solid and Illumina/Solexa, which differ from each other for the amount of data sequenced to each experiment. In this case, TopHat receives as input reads produced by the Illumina Genome Analyzer, although users have been successful in using TopHat with reads from other technologies. The input samples consist of two files of about 37 million of reads each. The two files are FASTA formatted paired-end reads. Dealing with paired-end reads means that the reads are sequenced by the sequencing machine only on the end of the same DNA/RNA molecule, thus the sequence in the middle part is unknown. Each sequenced end of the same read is also referred as mate. It results in two distinct files: the first mate of the same reads and the opposite mate. TopHat finds junctions by mapping reads

to the reference in two phases. In the first phase, the pipeline maps all reads to the reference genome using Bowtie. All reads that do not map to the genome are set aside as initially unmapped reads. Bowtie reports, for each read, one or more alignment containing no more than a few mismatches in the 5'-most s bases of the read. The remaining portion of the read on the 3' end may have additional mismatches, provided that the Phred-quality-weighted Hamming distance is less than a specified threshold. TopHat allows Bowtie to report more than one alignment for a read, and suppresses all alignments for reads that have more than this number. This policy allows so called multireads from genes with multiple copies to be reported, but excludes alignments to low-complexity sequence, to which failed reads often align and then assembles the mapped reads. TopHat extracts the sequences for the resulting islands of contiguous sequence from the sparse consensus, inferring them to be putative exons. TopHat produces a compact consensus file containing called bases and the corresponding reference bases in order to generate the island sequences. TopHat uses the reference genome to call the base. Because most reads covering the ends of exons will also span splice junctions, the ends of exons in the pseudoconsensus will initially be covered by few reads, and as a result, an exons pseudoconsensus will likely be missing a small amount of sequence on each end. In order to capture this sequence along with donor and acceptor sites from flanking introns, TopHat includes a small amount of flanking sequence from the reference on both sides of each island. TopHat has a number of parameters and options, and their default values are tuned for processing mammalian RNA-Seq reads also it can be used for another class of organism.

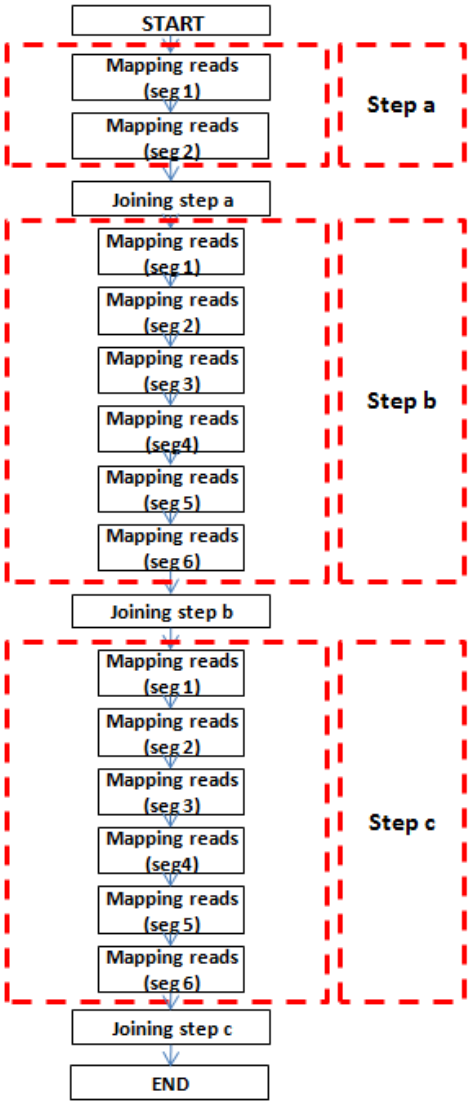


Figure 2. Sequential TopHat Workflow.

B. Alignment Tools: Bowtie

The short reads alignment is surely the most common operation in RNA-Seq data analysis [7]. The purpose of the alignment is to map each short read fragment onto a genome reference (see Figure 1). From the computational point of view, each short read consists of a sequence of four possible characters corresponding to the DNA bases and the sequence length depends on the sequencing machine adopted for the biological experiment [8]. The main novelty introduced by NGS technology is the capability of sequencing small DNA/RNA fragments in parallel, increasing the throughput and producing very short reads as output. However, this feature make the computational problem more challenging because of the higher amount of read produced and the accuracy in the mapping (shorter sequence length, higher probability of having multiple matches). For this reason many alignment tools specifically focussed on the alignment of short reads have been recently developed. In the present work, we are interested in characterizing the performances of alignment tools on real NGS data. Given our context, TopHat has been chosen, a wide diffused alignment program particularly aimed at aligning short reads. In order to detect the actual limitation of the alignment phase, we considered real NGS data coming from the analysis of Chronic Myeloid Leukemia. In our analysis flow, the HG19 assembly produced in 2007 is considered as reference genome the last human genome assembly produced to now. In order to increase the computational performances during the read mapping, Bowtie program creates an index of the provided human genome reference. This operation is particularly straightforward from the computational point of view, but it must be performed only one time for the human genome reference and it is independent on the mapping samples. The alignment phase itself is particularly suitable to be parallelized. In fact, each mapping operation is applied to each read independently on the other read mapping.

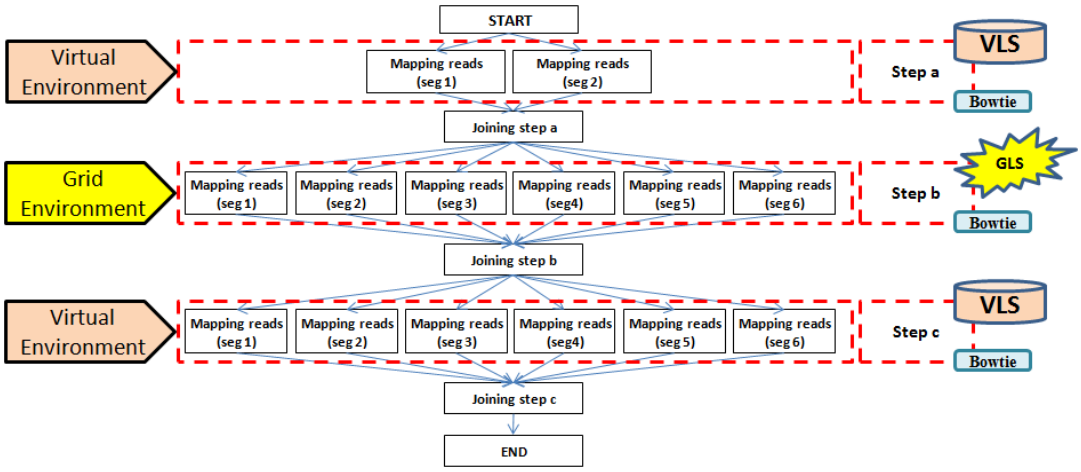


Figure 3. Distributed TopHat Workflow.

IV. Related work

Myrna [9] is a cloud computing tool for calculating differential gene expression in large RNASeq datasets. It allows to integrate short read alignment with interval calculations, normalization, aggregation and statistical modeling in a single computational pipeline. After alignment, Myrna calculates coverage for exons, genes, or coding regions and differential expression using either parametric or non-parametric permutation tests. The main advantage is the ability to rapidly test multiple plausible models for RNA-Seq differential expression. It has been suggested that this type of flexibility is necessary for computational applications to keep pace with the rapidly increasing number of reads in NGS data sets. Myrna allows to show that biological replicates reflect substantially increased variation compared to technical replicates in RNA-Seq and demonstrate that the commonly used Poisson model is not appropriate for biological replicates. It is designed to be run on the cloud using Amazon Elastic MapReduce, on any Hadoop cluster, or on a single computer. MapReduce is a programming model for processing large data sets implemented by Google. Usually, MapReduce is used to do distributed computing on clusters of computers. The model is inspired by the map and reduce functions commonly used in functional programming. MapReduce is a framework for processing parallel problems across huge datasets using a large number of node as cluster (Virtual Environment, Grid and Cloud Computing). Computational processing can occur on data stored either in a unstructured filesystem or in a structure database (also distributed DB). Algorithm is composed of two steps: Map step consists of divides task into smaller tasks, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller tasks, and passes the answer back to its master node. Reduce step: consists of collects the answers to all the sub-problems received from worker nodes and combines them in some way to form the output. MapReduce allows for distributed processing of the map and reduction operations. Main features is that each mapping operation must be independent of the others, so all maps can be performed in parallel. Hadoop is an open source software framework derived from Google's Map Reduce and Google File System for reliable, scalable, distributed computing of huge amounts of data. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. Hadoop allows to split them into smaller sub-tasks, before reducing the results into one master calculation. This technique is not new, it has been conceived with the grid computing that has a new life with the coming of cloud computing. AWS (Amazon Web Services) offers a set of infrastructure and application services that allow the users to run virtually enterprise applications, big data projects, social games and mobile apps. Since 2009, Amazon launched Elastic Map Reduce which provides a hosted and scalable Hadoop service. Moreover Amazon offers other big data related services such as the Simple Queue Service for coordinating distributed computing and RDS a Managed Relational Database Service. Elastic Map Reduce (EMR) is the Amazon's implementation of Hadoop, it can be programmed in a usual way with tools like Pig and Hive. Instead of HDFS (Hadoop Distributed File System), as main data source, Amazon can be used as Simple Storage Service (Amazon S3) in order to get data in and out. But when using structured data S3 can be unwieldy, for this reason Amazon offers DynamoDB which is the No SQL storage solution to cover functionality of Hadoop HBase. Amazon

offers also EC2 (Elastic Compute Cloud) [11] which is a web service that provides resizable compute capacity in the cloud. It offers a web service through which users can boot an Amazon Machine Image (AMI) to create a virtual machine ”instance” and rent it. Amazon offers pay-per-use services, then resources are paid in a manner based on compute capacity of an instance, the size of storage requests (GB) and network traffic generated. Elastic IP addresses are static IP (IPv4) designed for dynamic cloud computing. An Elastic IP address belongs to the account and not to a virtual machine instance. It exists until it is explicitly released by the user. Amazon EC2 is based on the XEN para-virtualization technology and it is possible to sizes instances based on EC2 Compute Unit (ECU). One EC2 Compute Unit provides the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor. Therefore the features of the virtual machine (such as the number of CPU cores, the processing power per core, memory, performance I/O, etc.) are fully user configurable and it is possible to choose from a series of configurations of the server. Even the operating system and the software that runs on the virtual server can be completely customized by the user.

V. METHODOLOGY

The proposed solution provides novelty on two aspects, firstly an optimization of the read mapping algorithm has been designed, in order to parallelize processes, secondly an implementation of a Hybrid architecture which consists of a Grid platform, composed of physical nodes, a Virtual platform, composed of virtual nodes set up on demand, and a scheduler devoted to integrate the two platforms.

A. Reverse Engineering

In a preliminary phase of reverse engineering, studying TopHat, blocks of transactions have been highlighted that were executed sequentially. We identified three main blocks, that can be executed independently:

a left and right mate mapped with HG19;
b segments mapped with HG19;
c segments mapped with segment juncs.

A feature of these three blocks is that they are performed by an external software, called Bowtie, as explained before. In steps (a) and (c), since the files involved in the development are significant, a common repository has been created that contains the temporary folders used by TopHat. Although, the use of a common repository slightly increases processing time, due to the SSH protocol connection, this time is lower than the time of transfer of the entire set of files to other machines. Instead the step (b) uses small files these can be performed on Grid environment, both physical and virtual, because the transfer times are lower. The only difference is that the input files are transferred to worker nodes through Globus Toolkit. These worker nodes re-send the output file to the node that requested execution when the process is terminated.

B. VirtualBio Infrastructure

The proposed architecture allows to manage RNA data, prepared by the version of TopHat in Grid, but not only, it could also handle other processing flows that use software and other tools, e.g., original version of TopHat or only Bowtie (see Figure 4). The architecture, called VirtualBio, is composed of three main components: a Master Node (MN), a part consists of the Physical Worker Nodes (PWN) that set the grid environment while a part consists of Virtual Worker Nodes (VWN) that set the virtualized environment. The MN is a physical machine with quite good hardware characteristics, is responsible for CA, contains the database, where all information about the nodes belonging to the infrastructure are stored, the node status, the flow of the various biological analysis that can be made in the system and system monitoring. Moreover, on it has been configured the common repository, using the Network File System (NFS). NFS is a protocol developed by Sun Microsystems in 1984 to allow computers to share files and folders over a network [13]. NFS is an open standard, defined in RFCs, and allows anyone to implement the protocol. Both environments are configured with the middleware Globus Toolkit [14], since it allows obtaining a reliable information technology infrastructure that enables the integrated, collaborative use of computers, networks, and databases. The Globus Toolkit is a collection of software components designed to support the development of applications for high performance distributed computing environments, or computational grids. OpenNebula.org [15] is an open-source project for building and managing virtualized enterprise data centers and cloud infrastructures. OpenNebula organizes existing storage, networking, virtualization, monitoring, and security platforms to enable the dynamic placement of multi-tier services. OpenNebula offers:

- The Image Repository System: it allows to set up and share images (e.g. operating systems or data) to be used in Virtual Machines easily.
- The Template Repository System: it allows to register Virtual Machine definitions in the system, to be instantiated later as Virtual Machine instances.
- Virtual Networking to interconnect Virtual Machines, they can be defined as fixed or ranged networks.
- As soon as a Template is instantiated to a Virtual Machine, can be performed a number of operations to control their lifecycle (migration, stop, resume, cancel, etc.). These operations are available both from the CLI and the Sunstone GUI.

Open Nebula corresponds most to the definition of a hybrid cloud. It integrates external resources in the cloud and does not introduce its own proprietary infrastructure. It has a

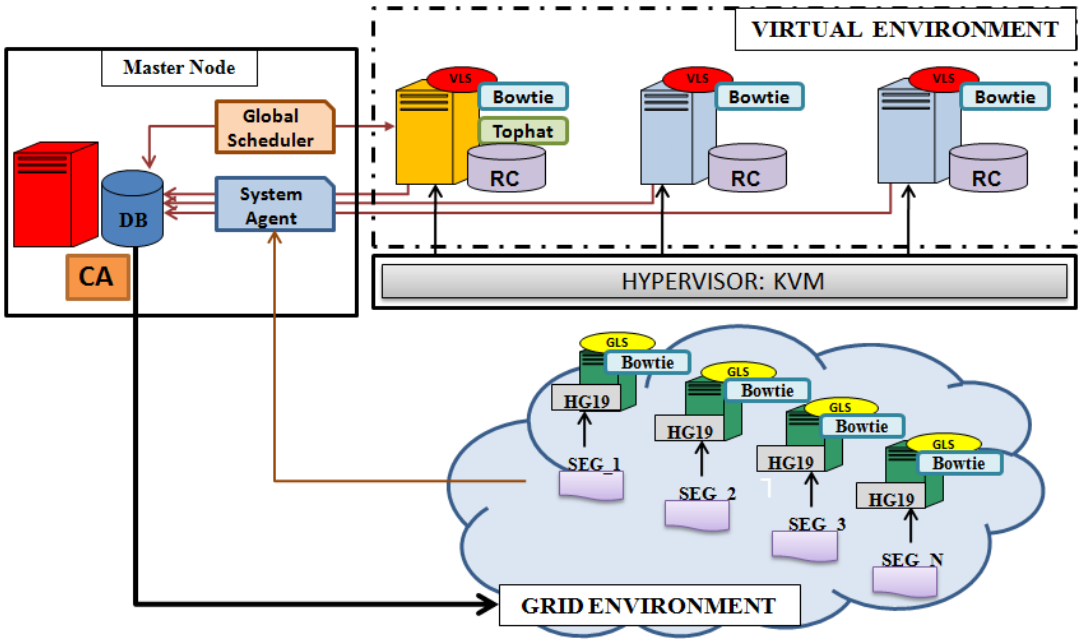


Figure 4. VirtualBio Architecture.

flexible component based structure and includes several pre-defined drivers for information management (monitoring), image and storage management (e.g., via NFS, LVM, or SSH), as well as to support several hypervisors (KVM, Xen, VMware). OpenNebula also integrates drivers for Amazon EC2 and ElasticHosts, and can be easily extended to support other cloud providers. With its support for EC2, OpenNebula can be configured to use Amazons infrastructure for cloud bursting, or use other EC2 compatible systems (such as Eucalyptus or Nimbus). An OpenNebula cloud typically consists of a front-end node for administration purposes (i.e., for managing hosts and images), as well as of several cluster nodes to execute the VM images. The hosts are controlled by the front-end either via command-line tools, or via well-defined programming interfaces.

1) Grid Environment: The Grid environment consists of machines with high computing power, this allows to use own machines and also machines belonging to different virtual organization. The only requirement is to have the necessary software installed for the processing (Bowtie, TopHat and Globus Toolkit). On each worker node of the grid environment is installed the Grid Local Scheduler, an essential component for performing biological tests.

2) Grid Local Scheduler: The Local Grid Scheduler (GLS) is a scheduler of active physical machines, which has been developed for the design phase (b), it aligns the segment with respect to the human genome (HG19) through Bowtie. Since the transfer of the input file is not influential, the worker nodes do not need to be in the same subnet as the Master Node, but may also belong to different virtual organization, so system can have greater scalability and can use machines powerful performance [10].

3) Virtual Environment: Virtualized environment also helps to improve infrastructure management, allowing the use of virtual node template to create virtual nodes in a short time, speeding up the integration of new nodes on the grid and, therefore, improving the reactivity and the scalability of the infrastructure [12]. The open source KVM has been used as hypervisor. It allows to create Fully Virtualized machines. The kernel component of KVM is included in mainline Linux [16]. KVM allows a Full Virtualization solution for Linux on x86 hardware containing virtualization extensions (Intel VT or AMD-V). KVM is implemented as a module within the Linux kernel. A hypervisor hosts the virtual machine images as regular Linux processes, so that each virtual machine image can use all of the features of the Linux kernel, including hardware, security, storage and applications. Full Virtualization provides emulation of the underlying platform on which a guest operating system and application set run without modifications and unaware that the platform is virtualized (see Figure 4). It implies that every platform device is emulated with enough details to permit the guest OS to manipulate them at their native level. Moreover, it allows administrators to create guests that use different operating systems. These guests have no knowledge about the host OS since they are not aware that the hardware they see is not real but emulated. The guests, however, require real computing resources from the host, so they use a hypervisor to coordinate instructions to the CPU. The main advantage of this paradigm concerns the ability to run

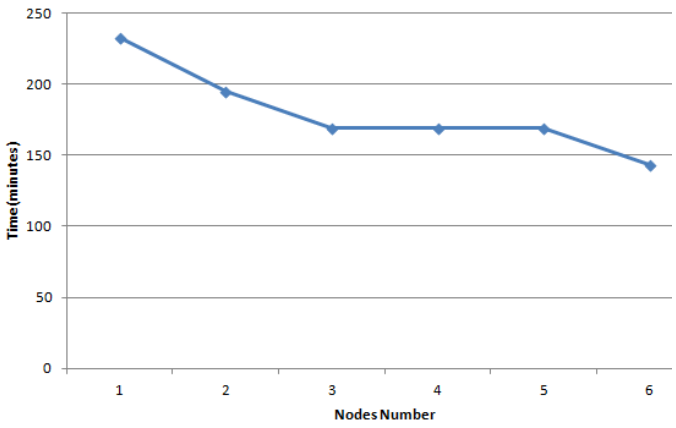


Figure 5. TopHat Execution Time for a Single Sample.

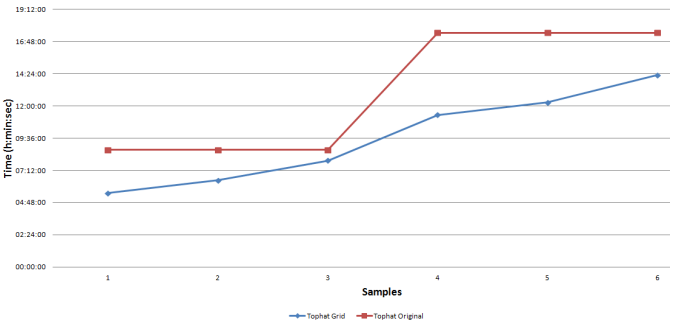


Figure 6. Original TopHat vs TopHat Grid.

virtual machines on all popular operating systems without requiring them to be modified since the emulated hardware is completely transparent. The virtualized environment has pre-installed images, which contain all software and libraries needed for running Bowtie and TopHat. The pre-configured images allow an ease instantiation of the machines when needed, and can be easily shutdown after the use.

4) Virtual Local Scheduler: The Virtual Local Scheduler (VLS) is a scheduler of active virtual machines. Its purpose is to draw up the steps (a) and (c) of TopHat. As the GLS, the VLS performs the mapping files for input received through Bowtie. The step (a) allows the alignment with respect to the human genome (HG19) and step (c) allows the alignment with respect to the fragment juncs previously constituted by TopHat. Since the considerable size of the files involved in these two steps, the VLS works directly on the temporary folder that is located in the common repository, allowing to avoid wasting time due to the transfer of data. Even in this case the interaction with the database is essential and very frequent, network problems may affect the entire biological analysis.

VI. Performance Considerations

In a preliminary work, we have introduced two case studies, based on Bowtie execution, from two different points of view. The first is the fragment size, while the other one is the CPUs number on the worker node. In Table I a summary of the calculations obtained changing CPUs number are presented. We can notice for reads between 100 and 1000000 no gain of time has occurred, so for our studies only reads from 1000000 to 85000000 are considered. This is because the files have limited data, thus the processing times are already reduced at this stage and then having multiple processors is irrelevant. As we explained before, during an analysis phase of the algorithm, 3 main blocks have been identified, (a) left and right mate aligned with HG19, (b) segments aligned with HG19, (c) segments aligned with segment juncs. The processing time of each segment depends on parameter pthread that is specified in command of Bowtie and refers to the number of parallel processes that can run. Figure 5 depicts the processing time of entire flow of TopHat for a single sample when nodes number increases. Elaboration with only a node corresponds to original version of TopHat, it means in sequential version. We want to focus the attention on elaboration time when 3/4/5 nodes are available, we obtained no gain of time because each node has more than one segment to process.

For this test phase, we wanted to use an architecture which consists of three machines with four CPUs. In Figure 6, we can notice that already only a sample processed with the version of TopHat Grid, a time savings of 40% is obtained instead increasing the number of samples to be processed, it is worth noting that the percentage of earned time is about 30%, this is due to the jobs queues that are created on the nodes. In Figure 7, processing time of a single segment of the variation of the parameter pthread is depicted. The processing time of each segment depends on parameter pthread that is specified in command of Bowtie and refers to the number of parallel processes that can run. Once past this threshold, the trend is no longer regular, this is due to the scheduling allocation of the CPU operating system. This test allowed to have a vision on the processing time will have access to machines with different power, opening to a more accurate scheduling policy adapted to the needs of time of the biologist. The test was run on a machine with 12 CPUs, it is worth noting that in order to gain the maximum time the number of pthread must be equal to the number of CPUs.

In Figure 7, processing time of a single segment of the variation of the parameter pthread is depicted. The test was run on a machine with the following hardware characteristics: Intel Xeon CPU X5660 @ 2.80 GHz, 12 CPUs and 20 GB of RAM, it is worth noting that in order to gain the maximum time the number of pthread must be equal to the number of CPUs. Once past this threshold, the trend is no longer regular, this is due to the scheduling allocation of the CPU operating system. This test allowed to have a vision on the processing time will have access to machines

Table I
BOWTIE EXECUTION TIME (SECONDS)

Reads	1 CPU	2 CPU	3 CPU	4 CPU	5 CPU	6 CPU	7 CPU	8 CPU
100	31	24	24	24	24	25	32	26
1000	26	38	24	24	24	25	13	20
10000	24	24	22	23	22	23	22	22
100000	28	25	25	24	24	24	24	24
1 E06	62	49	50	42	41	41	38	38
10 E06	424	258	230	185	176	165	158	154
85 E06	3425	2044	1791	1432	1342	1247	1180	1048

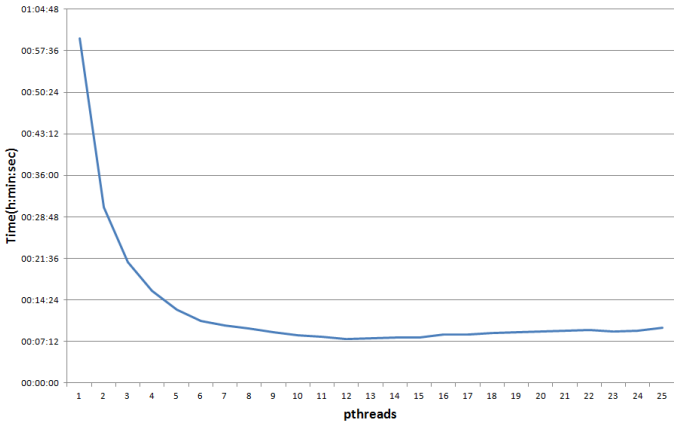


Figure 7. Bowtie Execution Time.

with different power and CPUs number, opening to a more accurate scheduling policy adapted to the needs of time of the biologist. Instead, Table I depicts the processing time of entire flow of TopHat, comparing the original algorithm (sequential version) with the modified algorithm (parallel version). We obtain a considerable gain of time that varies depending on the power of machines available. In our tests in order to make homogenous system, we have used machines with the same hardware (12 CPUs).

A. Amazon testbed

For the performed tests on Amazon EC2 was used a virtual machine High-Memory Quadruple Extra Large Instance (m2.4xlarge) [17] with the following characteristics:

- 68.4 GB of memory RAM;
- 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each);
- 100 GB of instance storage;
- 64-bit platform;
- Cost $2,28/hour.

This virtual machine has been added to the worker nodes and has been personalized installing Bowtie and all grid services necessary to enclose this node in the Virtual Grid. The executed tests on this machine show that although there is a decrease of the execution time (was used an 8 virtual cores instance) the transfer time increases however. Table II shows the time necessary to execute 6 fragments and the total transfer time. The input files for each fragment are 500 MB and the output files are 1.2 GB.

The usage costs for this instance were $6.84. It is noted that the transfer time is very high, almost half of the total time. To resolve this problem there are two possibilities: move all the Virtual Grid Infrastructure on Amazon, but the usage costs would increase or split the fragments to be processed further.

Table II
AMAZON'S INSTANCE USAGE TIME

Total Execution Time	Total Transfer Time INPUT Files	Total Transfer Time OUTPUT Files	Total Usage Time
01:12:00	00:24:00	00:48:00	2:24:00

VII. CONCLUSION AND FUTURE WORKS

VirtualBio is a platform for NGS analysis, with particular focus to the read alignment process using TopHat and Bowtie. The authors want to offer to biologist a private hybrid infrastructure to conduct their studies. After a careful study of existing solutions and state of the art such as Myrna, Supersplat and SplitSeek, we designed and developed the proposed solution. The novelty of this solution covers two aspects: computational infrastructure and optimization algorithm of TopHat. The hybrid infrastructure is composed of two environments: one based on grid computing and the other one virtual environment and with implementation of a common repository and a set of job schedulers. The second novelty covers algorithm aspects, during reverse engineering code of TopHat algorithm, the workflow has been optimized making parallel tasks that before were sequential. The proposed architecture allowed to reduce the elaboration time by at least 40% without additional cost due to the purchase of hardware. Future works include the optimization of scheduling policies opening the way to a multisample scenario and implementation of the architecture in Cloud environment, thus increasing the system scalability. It can include the integration of this solution in European Grid Infrastructure in order to able to submit tasks across the Europe

REFERENCES

[1] O. Terzo, L. Mossucca, A. Acquaviva, F. Abate and R. Provenzano, *Virtual Environment for Next Generation Sequencing Analysis*, The Eighth International Conference on Networking and Services (ICNS 2012), St. Maarten, Netherlands Antilles, 25-30 March 2012, IARIA, pp. 47-51

[2] De Magalhes J.P., Finch C.E. and Janssens G., *Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions.*, Ageing Research Reviews, 2010 Jul; Vol. 9(3), pp. 315-323

[3] Sanger F., Nicklen S. and Coulson A.R., *DNA sequencing with chain-terminating inhibitors*, Proc. Natl. Acad. Sci. USA 74, 1977), pp. 5463-5467

[4] Pop M. and S.L. Salzberg, *Bioinformatics challenges of new sequencing technology*, Trends Genet. Vol. 24, 2008

[5] Kircher M. and Kelso J., *High-throughput DNA sequencing concepts and limitations.*, Bioessays. 2010 Jun, Vol. 32(6), pp. 524-356

[6] Trapnell C., Pachter L. and Salzberg Steven L. *TopHat: discovering splice junctions with RNA-Seq*, Bioinformatics (2009), Vol. 25, Available at: doi:10.1093/bioinformatics/btp120, pp. 1105-1111

[7] Langmead B., Trapnell C., Pop M. and Salzberg Steven L. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology 10:R25

[8] Maher C.A., Palanisamy N., Brenner J.C., Cao X., Kalyana-Sundaram S., Luo S, Khrebtukova I., Barrette T.R., Grasso C., Yu J., Lonigro R.J., Schroth G., Kumar-Sinha C., Chinnaiyan Y., *Chimeric transcript discovery by paired-end transcriptome sequencing*, AM. Proc Natl Acad Sci USA, 2009 July, Vol. 28

[9] Langmead B., Hansen K. and Leek J. *Cloud-scale RNA-sequencing differential expression analysis with Myrna* Genome Biology (2010) 11:R83

[10] Kurowski K., Nabrzyski J., Oleksiak A. and Weglarz J. *Scheduling jobs on the Grid-Multicriteria approach*, Computational Methods in Science and Technology, 12(2), pp. 123-138, 2006

[11] Amazon Elastic Compute Cloud (Amazon EC2) Available at: http://aws.amazon.com/ec2/, January, 2012

[12] A. Chierici and R. Veraldi, *A quantitative comparison between xen and kvm*, Proc. of 17th International Conference on Computing in High Energu and Nuclear Physics, Journal of Physics: Conference Series 219 (2010).

[13] Network File System, http://wiki.ubuntu-it.org/Server/Nfs, December, 2011

[14] Globus Toolkit, Available at: http://www.globus.org/toolkit/, January, 2012

[15] OpenNebula, Available at: http://opennebula.org/, January, 2012

[16] Kernel-based Virtual Machine, Available at: http://www.linux-kvm.org/, December, 2011

[17] AWS Amazon, Available at: http://aws.amazon.com/ec2/instance-types/, January, 2012

A MANET Architecture for Airborne Networks with Directional Antennas

William Huba, Nirmala Shenoy
Golisano College of Computing and Information Sciences
Rochester Institute of Technology, Rochester NY 14623, USA
nxsvks@rit.edu, bill.huba@gmail.com

Abstract—**Surveillance using unmanned aerial vehicles (UAVs) is an important application in tactical networks. Such networks are challenged by frequent link and route breaks due to highly dynamic network topologies. This challenge can be addressed through robust routing algorithms and protocols. Depending on the surveillance area to be covered and the transmission range of the transmitters in the UAVs, several of them may have to be deployed, requiring solutions that are scalable. The use of directional antennas mitigates the challenges due to limited bandwidth, but requires a scheduling algorithm to provide conflict free schedules to transmitting nodes. In this article we introduce a new approach, which uses a single algorithm (i) that facilitates multi hop overlapped cluster formations to address scalability and data aggregation; (ii) provides robust multiple routes from data originating nodes to data aggregation node and (iii) aids in performing distributed scheduling using a Time Division Multiple Access (TDMA) protocol. The integrated solution was modeled in Opnet and evaluated for success rate in packet delivery and average end to end packet delivery latency. High success rates combined with low latencies in the proposed solution validates the use of the approach for surveillance applications.**

Keywords- Airborne Surveillance; Network of Unmanned Aerial Vehicles; Directional Antennas; Time Division Multiple Access; Distributed Scehduling

I. INTRODUCTION

Surveillance networks comprising of airborne nodes such as unmanned aerial vehicles (UAVs) are a category of mobile ad hoc networks (MANETs), where nodes are travelling at speeds of 300 to 400 Kmph. Surveillance requires aggregation of data captured by all nodes in the network at few nodes, from where the data is then sent to a center for further action. Due to high mobility of nodes and varying wireless environment, the topology in surveillance networks is subject to frequent and sporadic changes. Such MANETs thus face severe challenges when forwarding data from node to node, which is the task of the medium access control (MAC) protocol and also in discovering and maintaining routes between source and destination nodes, which is the task of the routing protocols. Another challenge faced is the scalability of the protocols to increasing number of nodes.

In this article, a unique solution for surveillance networks comprising of UAVs, equipped with directional antennas is proposed and investigated. The solution *uses a single algorithm* for several operations such as (i) multi-hop overlapped cluster formation, (ii) routing of data from cluster clients to cluster head for data aggregation, and (iii) scheduling concurrent time slots to transmitting nodes using a Time Division Multiple Access (TDMA) based MAC protocol, in UAVs that use directional antenna systems. To best leverage the strengths of this approach, the MAC, clustering and routing functions were implemented as processes operating using a single address generated by the algorithm to collaboratively address the challenges faced in surveillance networks. This leads to a new MANET architecture. Due to the critical nature of the application the new architecture and a unified approach is justified. Performance evaluations conducted in airborne networks with twenty, fifty and seventy five UAVs validate these justifications.

Surveillance applications require low packet loss and low packet delivery latencies hence in this work the analysis was directed primarily towards these performance metrics. Other performance metrics such as MAC and routing operational overhead were also recorded. The architecture introduced in this article achieves the performance goals. Due to lack of similar published work and the availability of evaluation models of implementations in such application scenarios, the presentation in this article is limited to the results from simulations of the proposed architecture.

The rest of the paper is organized as follows. Section II describes related work in the area of TDMA MAC, routing in large MANETs and clustering. The benefits of the integrated approach are highlighted in the light of these discussions. Section III describes the Integration Architecture, the rationale for the same, the components of the architecture and the interworking principles. Section IV describes the scheduler and the link assignment strategy. Section V provides the simulation details in Opnet and the performance analysis based on data collected. Conclusions and possible enhancements are discussed in Section VI.

II. RELATED WORK

The topic areas of major contribution in this article relate to routing, clustering and medium access control for use with directional antennas in MANETs. The significance of the proposed solution lies in the closely integrated operations of routing, clustering and medium access control coordinated by a control entity which has intelligence to coordinate their operations based on the applications requirements. To the best of our knowledge integrated clustering, MAC and routing solutions to MANETs have not been investigated though integration of clustering and routing have been researched. One of the main goals in this approach was to break down the limitation of protocol

layering towards an efficient MANET solution. Cross layered approaches, which break down such limitations in inter-layer communications, also facilitate a more effective integration and coordination between protocol layers. However, the proposed solution is not a cross layered approach, as one main problem encountered in such approaches is still the integration framework that has to work across different techniques and algorithm used by the different routing, clustering and MAC protocols. If the MAC uses a scheduled TDMA approach that has to work with directional antennas, the challenges are compounded. It was felt that for dedicated and critical MANET applications, one should not be constrained by the protocol layers or stacks, and other existing norms in the regard. What is important though is that such solutions should co-exist and interwork with networks that use current protocol structures.

Due to the uniqueness of the approach, it is not possible to cite and discuss related work that adopts similar techniques. Hence related work in each of the component topic areas are discussed under several subsections. Subsection A describes related work in the area of directional antennas and scheduled MAC protocols. This is followed by routing protocols and algorithms to address scalability in Subsection B. Hierarchical and hybrid routing protocols fall under this category. Routing combined with clustering is another approach to address scalability in MANETs and are discussed in Subsection C. Clustering, especially multi-hop clustering is very important to address scalability in MANETs; they also aid in data aggregation which is important in surveillance applications and is discussed in Subsection D. Lastly in Subsection E the significance of the proposed approach in the light of the related work is discussed.

A. Scheduled MAC in Directional Antenna Systems

To achieve higher capacity and improved delay guarantees in the network, Spatial reuse TDMA (STDMA) scheme is employed at the MAC layer [1, 2, 3]. In STDMA, which is an extension of TDMA, time is divided into time slots; and multiple transmissions can be scheduled as long as the receiving nodes do not get their packets interfered with. In this manner, STDMA takes advantage of the spatial separation between nodes to reuse the time slots. Generally, such schemes require strict time synchronization among participating nodes for efficient transmission and reception among the nodes. In addition, as a result of mobility of nodes in MANETs, periodic changes in the network require that STDMA schedules, which describe transmission rights of nodes in the network, be updated with minimal computational complexity. Furthermore, the updated schedule must be propagated to all nodes in the network in timely and efficient (using less resources) manner.

One of the most challenging tasks in such schemes is generating the STDMA schedule(s) that efficiently use the network resources. Since multiple nodes can simultaneously transmit in the same time slot, an optimal STDMA scheduling algorithm must allow high reuse of time slots with minimal interference while minimizing frame length (i.e. number of time slots per frame). Multiple algorithms have been proposed in literature [4-10]. The scheduling function can be performed by one of the participating node –a centralized scheduler. Centralized scheduling requires *all* information about the network such the number of nodes and links at the central scheduler, which is difficult to achieve. On the other hand, distributed scheduling can be done at the expense of increased complexity. In distributed scheduling, only nodes in the region of the change will act on it and update their schedules on network changes. In cluster based solutions centralized STDMA scheduling [6, 7], is less complicated and more efficient since each cluster head has all information about nodes in *its* cluster. Unfortunately, the overhead costs due to re-distribution of schedule whenever the network changes, are higher than that of distributed STDMA scheduling.

B. Routing in MANETs

Literature is rich with work conducted in the area of routing and clustering for MANETs. Several survey articles published on MANET routing and clustering schemes from different perspectives indicate the continuing challenges in this topic area. In [16] the authors present a survey of routing protocols and cross layer design effects. The survey presented in [17] is under the three broad categories of proactive, reactive and hybrid routing. A comprehensive technical report on MANET routing protocols [18, 20] covers them under the categories of uniform and non-uniform routing protocols, hierarchical (topology and cluster based), position based and so on, with performance comparisons. Reference [22] is an early review article that covers the characteristics of several routing protocols.

1) Proactive Routing Protocols

Proactive routing protocols require dissemination of link information periodically so that a node can use standard algorithms such as Dijkstra's to compute routes, to all other nodes in the network or in a given zone [27]. Link information dissemination requires flooding of messages that contain link information. Depending on the node mobility and wireless media conditions and the periodicity in link information dissemination, in large networks, such transmissions can consume significant amount of bandwidth making the proactive routing approach not scalable. Several proactive routing protocols thus target mechanisms to reduce this control overhead. *Fisheye State Routing* (FSR) introduces multi-level fisheye scope with reduced routing packet sizes and update frequency [28] to remote nodes. *Fuzzy Sighted Link State* uses the optimal routing algorithm, *Hazy Sighted Link State* [30] to further reduce link message dissemination. Multi scope approaches work well when the network grows in terms of number of hops end-to-end. *Optimized Link State* [25] reduces flooding of messages by using selected one hop nodes as *multi point relays*, to propagate link messages. *Topology Broadcast Reverse Path*

forwarding [29] propagates link-state updates in the reverse direction on a spanning tree formed by the minimum-hop paths from all nodes to the source node. The last two schemes achieve high efficiency in a dense network.

2) Reactive Routing Protocols

Reactive routing protocols avoid the periodic link information dissemination and allow a node to discover routes to a destination node only when it has data to send to that destination node. The reactive route discovery process can result in the source node receiving several route responses which it may cache. Routing overheads in reactive routing protocols can thus be considerably low if the number of simultaneously communicating nodes is not high. As mobility increases, route caching may become ineffective as pre-discovered routes may become stale and unusable. *Dynamic Source Routing* (DSR) [24] protocol, after the discovery, requires each data packet to carry the full address of every hop in the route, from source to the destination, and hence faces scalability problems as the addresses could be MAC (48 bits) or IP (32 bits) or IPv6 (128 bits). *Ad Hoc On-demand Distance Vector* (AODV) [23] routing protocol overcomes this problem by using intermediate nodes to maintain the forwarding information. *Temporally Ordered Routing Algorithm* (TORA) [18], [19]protocol uses link reversal, route repair and creation of *Directed Acyclic Graphs* (DAGs), similar to *Light-Weight Mobile Routing* (LMR) [34] and inheriting its benefits but reducing far-reaching control messages.

3) Hierarchical Routing

Partitioning a MANET physically or logically and introducing hierarchy can limit message flooding and also address the scalability. *Mobile Backbone Networks* (MBNs) [35] use hierarchy to form a higher level backbone network by utilizing special *backbone nodes* with low mobility to have an additional powerful radio to establish wireless link among them. LANMAR [34] was extended to route in the MBN.

4) Hybrid Routing

Scalability in MANET routing protocols have been addressed by combining proactive and reactive routing in a hybrid approach, where the use of proactive routing is restricted to a limited area or zone and reactive routing is used when communicating with distant nodes. Zoning requires some form of partitioning mechanism. *Sharp Hybrid Adaptive Routing Protocol* (SHARP) [36] is application adaptive and automatically finds the balance point between proactive and reactive routing. In SHARP, a hot destination node that receives data from many sources determines a proactive zone, and outside of the zone any reactive routing algorithm like AODV or DSR could be used. *Hybrid Routing for Path Optimality (HRPO)* [32] combines proactive route optimization to a reactive *source* routing protocol to reduce average end-to-end delay in packet transmissions. The *Zone Routing Protocol* (ZRP) [[33]] is a hybrid routing protocol, where each node has a pre-defined zone centered at itself. Any proactive routing can be used within the zone and any on-demand routing can be used for inter zone communications. ZRP provides a route discovery mechanism outside the zone through a *Bordercast Resolution Protocol* (BRP), where BRP establishes a *Bordercast* tree to send the discovery messages to the border nodes in a given zone.

C. Routing and Clustering

Nodes physically close to each other form clusters with a cluster head communicating on behalf of the cluster. *Multi Hop* clustering techniques such as the d-hop or k-hop clustering [8] algorithms can offer flexibility in terms of controlling the cluster size and cluster diameter, but are often complex to implement.

1) Clustering and Zoning

Clustering or zoning can be efficiently employed for the type of convergecast traffic encountered in surveillance networks, were the primary traffic flow is from cluster clients (CC) to cluster head (CH) [11- 15]. In such cases proactive routing approaches are recommended as the routing is limited to the cluster or zone and will also reduce stale routes. However proactive routing algorithms require the dissemination of link state information to all routers in the network or zone, which can introduce latency in realizing or breaking a route, and high overhead.

2) Cluster Based Routing

Different routing strategies can be used inside and outside the cluster. Several cluster based routing were designed to address scalability in MANETs. *Cluster Head Gateway Switch Routing* (CGSR) [15] is a cluster based hierarchical routing scheme. A mobile node belonging to two or more clusters acts as a *gateway* connecting the clusters. CGSR uses *distance vector* routing and maintains a cluster member table and a routing table at each node. *Hierarchical State Routing* (HSR) [16] is a multi-level, clustering based link state routing protocol that uses the clustering scheme recursively. In HSR, *Hierarchical ID* (HID) is used which is a sequence of MAC addresses of nodes on the path from the top of the hierarchy to the node.

D. Significance of the Architecture

From the above discussions it would be clear that clustering, routing and scheduling are different operations and hence normally are based on different algorithms or techniques. When combining the different operations, it becomes essential to define an interworking mechanism for the different algorithms. This adds processing complexity. It also results in added overhead for the operation of the combined functions. If all these operations can be based off a single algorithm, the complexity and overhead can be reduced significantly as demonstrated in this work.

If the above approach were possible, and if the MAC, routing, clustering and scheduling can use a single address

for their operation (unlike our current protocol stack, where MAC protocol uses 48 bit MAC addresses for its operation and routing protocols use 32 bit IP addresses (or 128 bits If IPv6)), we can achieve a solution, where the processes can closely interact and also avoid issues and overhead due to protocol layering, handling different headers and complex cross layered techniques. This would also make the solution compact and efficient and foster close interworking among the different operations.

III. The Integrated Approach

Given the challenges faced by MANETs, the authors decided to approach the solution from a holistic perspective. Towards this the essential functions required to support communications among the mobile entities in a MANET were identified. An architecture that would aid in best organizing the functions, taking into account the application demands and the challenging wireless media was then designed. The architecture would continue support for the existing protocol layered structure by either bypassing them during operation, interwork with them or replace them with a provision to bridge with networks based on these protocol structures.

The new architecture proposes a communications layer that bridges the application and the physical layers directly, bypassing other protocol layers. The communications layer includes routing, clustering and medium access functions whose operations are coordinated by an intelligent entity that incorporates the needs of the application taking into consideration the physical layer constraints.

A. The Rationale

Protocol layering introduces operational overhead. It also reduces efficient interworking among the protocols. Cross-layered techniques to address the communications needs of the wireless ad hoc networks were crafted for the purpose. Such techniques however introduce complexity as they overlay on the existing protocol structures. A few points to consider at this time is; (i) given a new wireless networking scenario and environment and the ensuing challenges, is there a need to continue with structures, algorithms and protocols that were developed for less challenging network situations such as the wired networks; (ii) secondly is there a need to continue with the two addresses in a bandwidth constrained environment? (iii) how about the complexity and resulting unreliability and lack of robustness?, and (iv) lastly how does this impact on the weight and power constraints faced by mobile devices? There is undoubtedly need for networks to interwork with one another, which does not however impose the condition that they have to use the same protocols, structures and so on.

B. The Architecture

A schematic of the architecture is shown in Figure 1. The light colored box indicates the use of either the TCP/IP protocol suite just below the Applications layer or the implementation of thin dummy protocol to incorporate port functions. The approach is similar to the Multiprotocol Label Switching used for tunneling to bypass IP layer often adopted in wired networks. It is different however as the communications layer now has all functions required for MANET operation, the MAC to enable sharing the wireless medium, the routing functions to discover routes reactively or proactively and clustering which is needed to address the scalability demands of MANET applications. All these functions are now coordinated by an intelligent Operation Control (OC) entity.

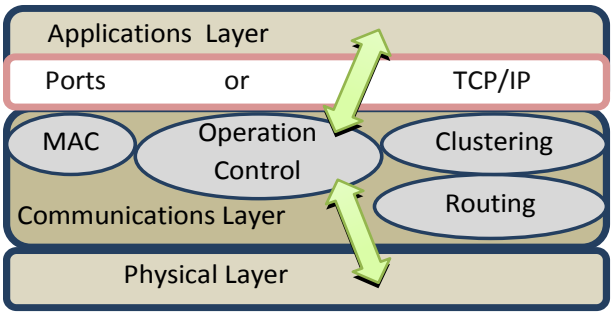


Figure 1 The Integration Architecture

The crucial entity in this architecture is the OC. Any MAC, routing or clustering protocol could be used in the other blocks. However, if these protocols operate on different techniques and address schemes, the effectiveness of the OC unit is reduced and it could become very complex balancing off the benefits of the approach. Note the positioning of the OC unit in the architecture (without TCP/IP suite) would provide information on the applications traffic and their quality requirements, whilst also collecting data on the Physical layer to control and coordinate the operations of the other entities in the communications layer. If TCP/IP were included then the information from the application can be passed through the Diffserv field in the IP (v4) header. However IP routing would be bypassed and the solution operates transparent to layer 3 protocols.

C. The Components

In this section, the different components used in the communications layer of the architecture in this work will be described. As the OC unit is crucial to the architecture, this will be the first component to be discussed. To make the OC unit efficient it is important to adopt an algorithm or technique that would allow coordinated operation of the other three entities in the communications layer. The significance of the coordinated operation would be clear at the end of this section and will be justified when the performance is discussed. The Multi-Meshed Tree (MMT) algorithm was selected for this purpose. This algorithm modified accordingly has already been used to support MAC, routing and clustering [37-41]. MMT is also amenable to optimization based on the network communications needs and is discussed under future work in the Conclusion Section. In this work, the algorithm was

enhanced to maintain several connections between nodes in a meshed tree cluster and the cluster head which is the root of the meshed tree. For completeness, the MMT algorithm is first briefly described. This is followed by clustering supported by MMT, the proactive route maintenance and then the establishment of reactive routes for communication among nodes between clusters. This is followed by the interworking principles between the scheduler, MAC and the directional antenna system.

1) Meshed Tree

It is a traditional approach to have mesh connections among communicating nodes for redundancy purposes, but the mesh is then logically configured (by blocking some of the physical connections) into a tree using either the spanning tree approach or the Dijkstra approach (or other tree

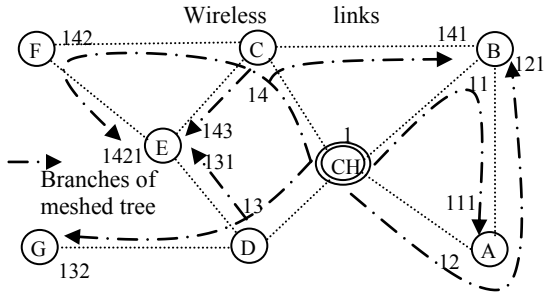


Figure 2 Meshed Trees

algorithms) to avoid looping packets. Logical tree creation adversely impacts the strengths of the meshed topology, as on any physical link changes the tree connections can break and the tree will have to be recreated. This is true of wired and wireless networks. In wireless ad hoc networks such link breaks can occur more often and the latency introduced in the network connectivity establishment and convergence can be detrimental.

Contrary to this approach, the meshed tree algorithm allows building several tree branches that exist concurrently from a single root by leveraging all the connections possible in the meshed structure without logically blocking any links. The tree branches so formed are limited only by criteria specified for the meshed tree creation. Looping is avoided through the use of a smart numbering scheme to define the tree branches.

Thus under the meshed tree algorithm a node resides in multiple tree branches unlike the trees formed by the Dijkstra or the spanning tree algorithms. On the failure of one path (tree branch) the node remains connected to the root on another path, without the need to rebuild the tree. Meshed tree construction is dynamic and the tree branches evolve continually based on the decisions by the nodes to join a branch. In the case of mobile nodes this feature allows the nodes to remain connected to the root with a high probability despite link breaks. Time lags and their impact to reconstruct the tree and their resultant performance impacts are avoided.

2) Meshed Tree Clusters

As per the proposed solution the meshed tree created around a designated or elected root is a cluster; the root node is the cluster head. The creation of a meshed tree is explained with the aid of Fig. 2. The dotted lines link nodes that are in communication range with one another at the physical layer. The root of the meshed tree is labeled 'CH' for cluster head. Nodes A to G are the cluster clients (CC). For simplicity in explanation, the meshed tree formation is restricted to nodes that are connected to the CH, by a maximum of 3 hops. At each node several *values or IDs* are noted. These are the virtual IDs (VIDs) assigned to the node as they join a meshed tree branch in the cluster. Let the CH be assigned a VID *1*, the CCs have *1* as a prefix in their VIDs. Any CC that attaches to a branch is assigned a VID, which inherits the prefix from its parent node, followed by an integer, which indicates the child number under that parent. In this work we limit the number of children to nine and use single digits to identify the children nodes. This does not eliminate the possibility of the scheme to have more than nine children under one node. It was not used in this case, as having too many paths going through a single node could create bottlenecks.

D. Routing in the Architecture

(i) Proactive Routes in the Cluster

In Fig. 2, each tree branch (shown by the dotted-dashed lines with an arrow head) is a sequence of VIDs that is assigned to CCs connecting at different points of the branch. The branch information of the meshed tree provides the route to send and receive data and control packets between the CCs and CH. For example, the branch denoted by VIDs 14, 142 and 1421, connects nodes C (via VID 14), F (via VID 142) and E (via VID 1421) respectively to the CH. To forward a packet from CH to node E, its VID 1421 will be used as the destination VID. When such a packet is broadcast, enroute nodes C and F receive the packet and forward to E. This is possible as the VIDs for nodes C and F are contained in E's VID. The VID of a node thus provides a virtual path vector from the CH to itself. Note that the CH could have also used VIDs 143 or 131 for node E, in which case the path taken by the packet would have been CH-C-E or CH-D-E respectively. Thus between the CH and node E there are multiple routes identified by the multiple VIDs. The support for multiple routes through the multiple VIDs, allows for robust and dynamic route adaptability to topology changes in the network and the cluster. Nodes can request for new VIDs and join different branches as their neighbors change.

To send a packet from node E to CH, the packet has to be directed to destination VID 1, which is its first digit. To send packets to other nodes in the cluster, the packet can be passed via the CH, a common parent node or to a child node

forward packet to other nodes either the cluster head will be used, or the packet can be sent directly on the branch if the source and destination node have (grand) parent or (grand)child relationship.

G. Scalability in the Arhcitecture

1) Inter-Cluster Overlap and Scalability

A surveillance network can comprise of several tens of nodes; hence the solutions for surveillance networks have to be scalable to that many nodes. We assume that several 'data aggregation nodes (i.e., CHs)' are uniformly distributed among the non-data aggregation nodes during deployment of the surveillance network. Meshed tree clusters can be formed around each of the data aggregation nodes by assuming them to be roots of the meshed trees. Nodes bordering two or more clusters are allowed to join the different meshed trees and thus reside in the branches originating from different CHs. Such border nodes will inform their CHs about their multiple VIDs under the different clusters. When a node moves away from one cluster, it can still be connected to other clusters, and thus the surveillance data collected by that node is not lost. Also, by allowing nodes to belong to multiple clusters, the single meshed tree cluster based data collection can be extended to multiple overlapping meshed tree (MMT) clusters that can collect data from several tens of nodes deployed over a wider area with a very low probability of losing any of the captured data. This addresses the scalability requirements in surveillance networks

Figure 3 shows 2 overlapped clusters and some border nodes that share multiple VIDs across the two clusters. The concept is extendable to several neighboring clusters. Nodes G and F have VIDs 142, 132 under CH1 and VIDs 251 and 252 under CH2, respectively.

2) Flexible Multi-hop Cluster Formation

Except for the CH, each node in Fig. 2 is a CC that will send the captured surveillance data to the CH. The size of the tree branch can be limited by limiting the length of the VID, which in turn allows control of the diameter of the cluster. Each node that joins the cluster has to register with the CH, by forwarding a registration request along the branch of the VID. This confirms the path defined by the VID and also allows the CH to accept /reject a joining node to control the cluster size. The number of VIDs allowed for a node can control the amount of meshing in the tree branches of the cluster.

Note that a node is aware of the cluster under which it has a VID as the information is inherent in the VIDs it acquires, thus a node has some intelligence to decide which VIDs it would like to acquire – i.e. it can decide to have several VIDs under one cluster, or acquire VIDs that span several clusters and so on. Moreover, a VID also contains information about number of hops it is from the CH, an attribute inherent in the VID length. This information can be used by a node to decide the cluster branch it would like to join based on the hops.

H. Inter-cluster Reactive Routing

This feature though not used in the work is described for completeness of the proposed architecture and its capabilities. Nodes bordering two or more clusters are allowed to join the branches originating from different CHs, and will accordingly inform their respective CHs about their multiple VIDs under the different clusters.

A node that has to discover a route to a distant node sends a 'route request' message to its CH(s). The CH then identifies the neighboring clusters based on updates from border nodes and forwards a copy of the 'route request' message to the border node, so that they can forward to the CH in the next cluster. The 'route request' message however has an entry for all the clusters that will be receiving the message, to avoid looping of the message. Thus the route request is not forwarded by all nodes, but only by all clusters and follows a path CH-border node- CH and so on.

When the CH of the destination node receives the route request, it will forward the route request directly to the destination node. The clusters forwarding the route request record the original sending node and the last cluster that the route request came from; this information is useful in forwarding the route response message when it returns. The destination node generates the route response and sends to its CH, which then forwards it back to the CH in the originating cluster and the source node along the same *cluster path* the route request took. Along the path back, all forwarding CHs will record the previous cluster and original sender of the route reply. The route between the sender and the destination node is thus initially set up as a sequence of CHs, but maintained as next cluster information. Mobility of

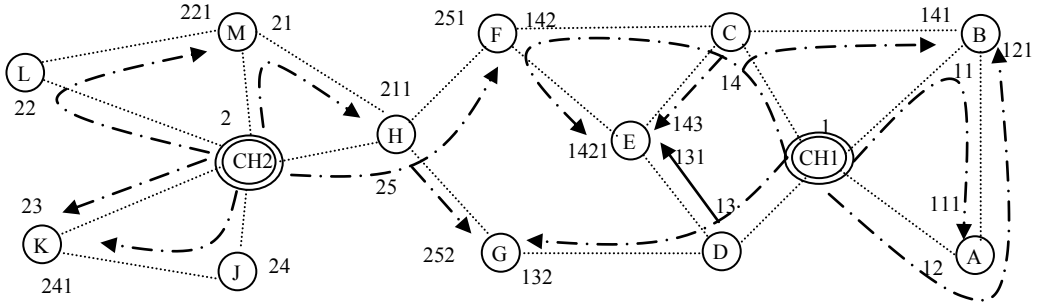


Figure 3 Overlapped Multiple Meshed Trees (MMT)

nodes does not impact the reactively discovered route, as long as the CHs exist. Note that movement of CHs also does not impact the reactive routes.

1) Robustness of the Reactive Routes

The route between nodes L in cluster 2 and A in cluster 1 while there are having an active sessions will be maintained at CH2 and CH1. If there were other clusters they would not maintain information for the route between the two nodes. Thus the reactively discovered route between L and A is maintained as a sequence of CHs and at the CHs as described earlier. The proactive route between L and CH2 and A and CH1 can change continually as the nodes move. Also the border nodes used between CH1 and CH2 to forward packet under the session can change, which change is recorded and maintained by the two CHs. Despite all the changes in the proactive routes, the reactive route which is the sequence CH2- CH1 does not change. They will change only if the CHs die. Thus the probability of a reactive route failure depends now on only two nodes as compared to the several numbers of nodes that normally define the reactively discovered path. With node mobility a single node movement in a path results in the path failure and rediscovery. In the proposed scheme as the reactive routes are a concatenation of the proactive routes between node-CH-border node-CH- node and these proactive routes are dynamically updated as the nodes move, reduces the probability of the reactive route failure considerably.

I. Highlights of the Architecture

Under the related work section we highlighted several routing schemes, and frameworks that combined different types of routing algorithms and cluster based routing. From the meshed tree based clustering and routing scheme described thus far, it should be clear that the scheme adopts a proactive routing approach, where the proactive routes between CCs and CH in a cluster are established as the meshed trees or clusters are formed around each cluster head. Thus a single algorithm and through process of joining a cluster nodes automatically also acquire routes to the CH. There is flexibility in dimensioning the cluster in terms of CC in a cluster and the maximum hops a CC is allowed from a CH. The tree formation is different from other tree algorithms as a node is allowed to simultaneously reside in several branches, and thus allowing for dynamic adaptability to route changes as nodes move. This also enhances robustness in connectivity to the CH. We know of no work in the literature with such unique properties. Though multiple overlapped clusters have been discussed in the literature [15], [16], the proposed meshed tree cluster achieves this in a simple way.

J. Interworking of Modules in the Architecture

It is important to understand the interworking of the modules and their interaction with the directional antenna system. Hence, the directional antenna system is first described followed by the interactions among the modules and their use of the directional antenna systems.

1) Directional Antenna System

All nodes in the surveillance network are assumed to be equipped with four phased array antennas capable of forming two beam widths. One beam width is focused with an angle of 10^o and the other is defocused with an angle of 90^o. The defocused beams are used for sending broadcast packets, while the focused beams are used for unicast or directed packets. Each antenna array covers a quadrant (90^o)

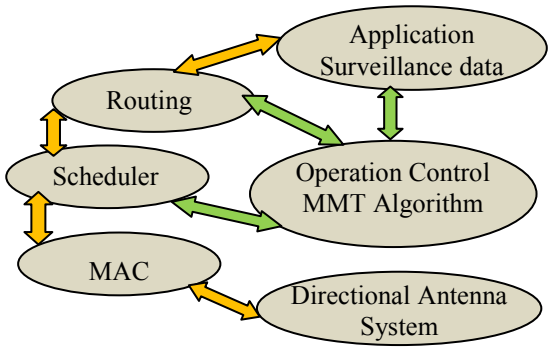


Figure 4 Interworking Modules

and is independently steerable to focus in a particular direction within that quadrant in the focused beam mode.

We also assume that each node is equipped with a Global Positioning System (GPS) which is used for time synchronization and to provide node position. The latter information is used in a tracking algorithm to estimate the location of a receiver node, so transmitting nodes can direct their beams to the destination node.

2) Interworking Principles

The surveillance data collected by the nodes is passed to the routing module, which will decide on the route or VID to use to forward the data to the CH based on directions provided by the OC. The OC unit in this case decided on routes with the least hops. When there is a backlog in the packet to a particular destination the OC unit informs the scheduler to negotiate for more slots. The meshed tree cluster formation and its parameters are maintained by the OC unit. The unit also decides on the overlap and number of VIDs to be maintained, the cluster size and so on. The OC unit can monitor Physical layer parameters to decide on the routes, this feature was not used in this work.

Once the route has been decided, the node knows the address of the next hop node which will forward the packet. This information is then passed to the STDMA scheduler to schedule slots, taking as input the number of slots, slot time and control slots. This information is then passed to the MAC to create the frame and forward to the next node. Before forwarding, the MAC, locates the destination node position and controls the antenna array to transmit the packet using a directed beam.

IV. SCHEDULING AND LINK ASSIGNMENT

The VIDs carry link information between a pair of nodes that share a parent-child relationship. Thus a link assignment strategy was adopted in this work. The structure of the VIDs, also allows each node in a cluster to be aware of its neighbors due to the parent-child relationship defined by the VIDs. This allows a node to schedule time slots with its neighbors (parent or child) taking into consideration its current committed time slots to its other neighbors.

A. Scheduler Operations

The scheduling algorithm has to schedule time slots for (1) cluster formation after deployment of the UAV nodes, (2) subsequent cluster and route maintenance, and (3) data aggregation. It should also send updated schedules in a timely manner as network topology changes. For all of these operations different categories of time slots as described below were used.

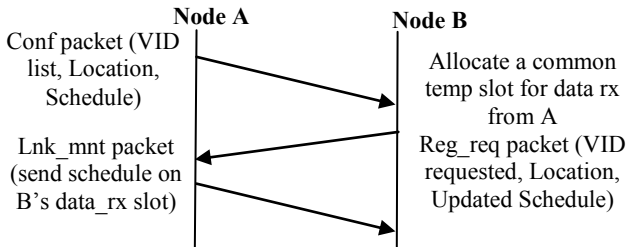


Figure 5. Distributed Scheduling Among Neighbors

Broadcast Slots: Some slots are preselected as broadcast slots in which they announce their VIDs, location, and current schedule, in a *configuration* (conf) packet, so neighboring nodes can listen and decide to join the cluster.

- *Directed Slots:* All other slots are used in a directed mode, where one node is transmitting using the directed beam to its listening neighbor. Directed slots can be *assigned* slots or *temp* (unassigned) slots.
- *Temp Slots* are used by nodes to negotiate for a common time slot for data transfer.
- *Assigned Slots:* Temp slots become assigned slots after a mutual negotiation by a pair of nodes. In the assigned slots control information for cluster and route maintenance, link maintenance (*lnk_ mnt*) control packet generated by the MAC and data packets are sent and received. Assigned slots are unidirectional and are used either for transmitting (data-tx) or receiving (data-rx). If there are data packets to be sent in such slots, the control packets are sent first, followed by the data packets. At least one packet must be sent by a node during in a data-tx slot each frame to every neighbor that it is associated with. When there are no data packets to send, the MAC sends lnk-mnt packets to monitor the link status. The link will be dropped between two nodes if these transmissions are not maintained every frame. Unidirectional links that can only send or receive data but not both are not supported in this scheme. Acknowledgement of received packets and retransmission of unacknowledged packets are handled by the MAC, but only route requests, route replies, and data packets are acknowledged. Lnk_mnt packets are implicitly acknowledged when the neighboring node sends its own lnk_mnt back. Each explicit ACK contains a low and a high sequence number which represent the range of packets than are being acknowledged. If the sender of the packet does not receive the corresponding ACK by the following frame, it will attempt to resend the packet up to a maximum of three attempts. At that point the link is considered failed and the VID is no longer valid. Any queued packets are rerouted after the VID is dropped.

B. Link Assignment

The approval of a new node by the CH is an indication that the CC has both a physical and logical path towards the CH. Scheduling slots for the new node starts subsequent to its acceptance into a cluster by the CH. Nodes individually schedule data slots in a distributed manner with their one-hop neighbors making the scheme truly distributed. The end to end information is carried by the VIDs. Time slots are scheduled for as long as at least one VID remains between a node pair. The process of mutual scheduling is explained with the aid of Fig. 3 below.

When node A advertises its VIDs via a *conf* packet it attaches its current schedule and GPS coordinates. Node B receives the packet and decides to request a VID under one of the advertised VIDs. Node B will then reserve a data-rx slot from one of the temp slots advertised by the parent that matches with its own temp slot and responds with a registration request, and the updated schedule to node A. Node A in turn assigns another temp slot that is common to the pair as a data-rx slot for receiving packets from node B.

Slot	1	2	3	4	5	6	7	8	9	10
CH	TX to B	RX to B	CTRL	TX to A	RX to A	TX to D	RX to D	TEMP	TX to C	RX to C
A	TEMP	CTRL	TX to B	RX to CH	TX to CH	RX to B	TEMP	TEMP	TEMP	TEMP
B	RX to CH	TX to CH	RX to A	TX to C	RX to C	TX to A	TEMP	CTRL	TEMP	TEMP
C	TX to E	TX to F	RX to E	RX to B	TX to B	CTRL	RX to F	TEMP	RX to CH	TX to CH
D	CTRL	TX to E	TX to G	RX to E	RX to G	RX to CH	TX to CH	TEMP	TEMP	TEMP
E	RX to C	RX to D	TX to C	TX to D	CTRL	TEMP	TEMP	RX to F	TX to F	TEMP
F	TEMP	RX to C	CTRL	TEMP	TEMP	TEMP	TX to C	TX to E	RX to E	CTEL
G	TEMP	TEMP	RX to D	TEMP	TX to G	TEMP	TEMP	TEMP	CTRL	TEMP

Table 1 Sample Schedule Generated by the Distributed Scheduler

It then forwards the registration request from node B towards the CH. During the next frame, node A will send a *lnk_mnt* packet to node B with the updated schedule. Thus a set of slots for transmitting and receiving between nodes A and B are decided. No other node's schedule is taken into account unless it directly affects the current link between two negotiating nodes. Time slots are reused across different sets of nodes by taking advantage of the spatial separation between nodes.

The process of allowing a new requesting node a VID to reserve a data_rx slot in which the parent node can transmit allows the parent node to resolve conflicts in case the suggested data_rx slot is not available. If two nodes attempt to assign themselves the same data-rx slot for a third node, the third node will accept the data-rx allocation from the first schedule that it receives. When it gets the second schedule, it will not make any schedule changes and just send a link maintenance packet to the sender. The denied node will see the conflict and choose a different temp slot to allocate as a data-rx slot. This also prevents any lost packets during the link establishment process. Since nodes choose their own receiving slots, but not transmit slots, there is certainty that the neighboring node is available during a transmission. Assigning data slots in this manner allows for dynamic asynchronous links.

For example if node A's buffer indicates packet (to be sent to node B) accumulation beyond a threshold value, then in the next lnk-mnt packet, A can request node B to set aside *x* data-rx slots, where the value x is capped to avoid one node taking up all available slots. Node B will respond with the updated schedule by setting aside the *x* slots provided it has no such similar demands from its other neighbors. If there are similar demands, it will allocate slots proportional to the demands of tis neighbors. The on demand allocation can result in increased number of data-rx slots at B (to receive from node A) though the single data-tx towards node A will be maintained unless changed by a demand. The tuning of the on-demand slots is executed every frame. If the amount of traffic being sent to node B decreases, the link will be reduced to having one data-rx and one data-tx slot again.

Table 1 is a sample schedule generated for the cluster in Fig 1. Nodes A, B, C, and D receive the initial configuration packet from the cluster head and schedule their data-rx (RX) slots; 4, 1, 9, and 6 respectively. This decision is a random allocation of matching temp slots based on the sequence in which the configuration packets were received. The cluster head accepts these data-rx packets which were sent in the registration request messages of these nodes and sets the corresponding slots as data-tx slots in its own schedule. It then allocates data-rx slots to each of these nodes on slots 5, 2, 10, and 7. Node A receives a configuration packet from Node B and decides to use slot 6 as its data-rx slot for receiving from node B. Node B then chooses slot 3 as the data-rx slot for A. At the same time Node B receives Node C's configuration and chooses slot 5 as its data-rx slot. Node C selects slot 4 as the complementary slot. This is the same slot that the CH is transmitting to Node A, but due to the directional antennas there will be no interference. The process continues branching outward until every link has a pair of slots allocated

V. SIMULATIONS

The performance evaluations of the surveillance network using the proposed solution was carried out using Opnet (version 14.5) simulation tool. All the processes explained above were modeled in Opnet. For surveillance data, each CC generated a 1 MByte file, which was then sent to the CH for aggregation. Normally UAVs travel in elliptical trajectories. In the models, we used circular orbits, to introduce more route breaks and thus stress test the solution. These circular orbits had a diameter of 20 Km (which defines the areas for each scenario), while the maximum transmission range was limited to 15 Km. the overlap between trajectories is seen in Fig. 4. A maximum of 5 UAVs were allowed in one circular trajectory, thus the UAVs were deployed over a wider area, which was covered with several trajectories. For example, in the 20 node scenario, there were four circular trajectories with slight overlap in their trajectories, to avoid physical network segmentation as shown in Figure 6. In the trajectories, the speed of the UAVs varied between 300 to 400 Kmph; hence, the different colors for the trajectories.

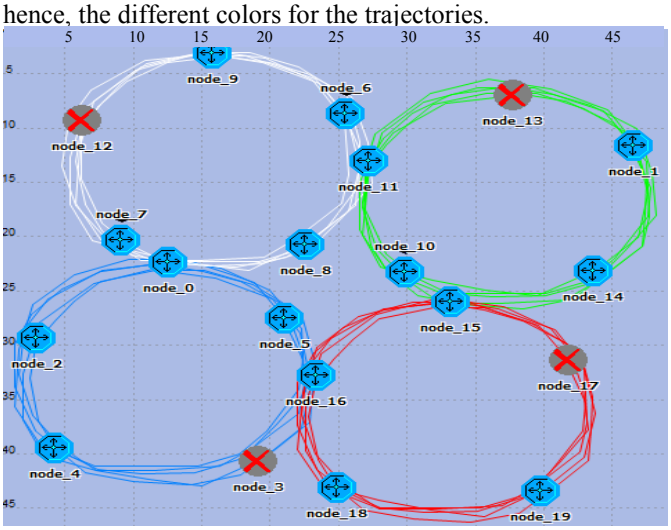


Figure 6. Typical Deployment and UAVs

The physical layer parameters were maintained invariant. Packets with 1 bit error rate were dropped and no *Forward Error Correction* was implemented. In the focused beam mode the data rate is 50 Mbps and in the defocused mode the data rate is 1.5 Mbps. A single frame had 50 timeslots each of 4 ms duration and 0.5 ms guard time. These values were optimized based on our prior work [4, 5].

Due to the lack of similar published work and models in Opnet (the evaluation tool used) the performance of the presented solution is analyzed with respect to the performance goals stated for surveillance networks earlier

namely success in packet delivery, and latency in packet and file deliveries. Included in the performance graphs are the overhead incurred by the MAC and routing protocols, and the average hops encountered during packet delivery, which is useful in explaining some results.

Success Rate was calculated as the percentage of packets received at a destination node with respect to the number of packets generated by the sender paired with that destination node.

Overhead for both MAC and routing was calculated as the percentage of control traffic to all the traffic in the network. This was determined only when data sessions were active. The bits contributing to overhead calculations was discussed earlier.

Packet latency was recorded as the end to end latency i.e. from the time the packet was sent by a sender node till it was received by the CH in seconds. File delivery latency was calculated similarly in seconds.

In each of the test scenarios, a certain number of nodes were randomly selected to send a 1 MByte file to the CH. These selected nodes sent the files simultaneously, thus stress testing the solution. Furthermore the number of sending nodes was increased to include all of the nodes except the data aggregation nodes, which is a highly stressful test scenario. Each test scenario was repeated with 20 different seeds (high prime numbers) and the results averaged over these seeds. The simulations were limited 20 runs in each case due to the stable outcomes noticed with different seeds.

A. 20 Nodes Scenario

Figures 7A to 7C are the plots for the twenty UAV scenario with 4 clusters. The x axis in all plots shows the number of nodes that are simultaneously sending aggregation traffic, i.e., 1 MByte file to the 4 CHs. The number of sending nodes was varied from 4 to 16. In the last case all 16 CCs were sending a 1 MByte file simultaneously to the CHs.

With increasing number of senders, the success rate hardly dropped below 100%. This shows the efficiency of the scheduler to successfully schedule all the packets that are arriving simultaneously. The average hops recorded in graph 1 however shows a decrease when the number of sending nodes was increased. When 20 nodes were selected to send traffic they encountered an average hop distance of 1.8 hops; which dropped to 1.4 hops when all 16 nodes were sending traffic. This is because of the random way in which the sending nodes were selected. The average hops graph can be interpreted thus – the first four nodes that were selected were farther away from the CHs, but as more nodes were randomly picked they were closer to the CH. The impact of this is noticeable in the packet and file latencies recorded in graph B, which shows a decrease with increasing number of senders.

In Fig. 7B, the average packet latency recorded was less than 0.8 seconds. Acceptability of packets arriving at this latency depends on the criticality of the surveillance application. If an upper limit was specified then that could be used as a cut off to drop packets arriving late. The file delivery latency is only slightly higher at around 1.2 seconds, which shows that all packets in the 1 MByte file were transported from the data collection node to the aggregation nodes, i.e., the CH within the time.

Fig. 7C is the plot of a very important parameter as it shows the channel bandwidth used by the control traffic both by the MMT based routing protocol as well as the MAC protocol. The MAC and routing overhead were recorded to show the ratio of messages used for control purposes by two operations.

The MMT routing overhead was below 20% while the MAC overhead reduced from 10% when there were 4 sending nodes to less than 5% when there were 16 sending nodes. It should be noted that the MMT routing traffic also includes the cluster formation control traffic.

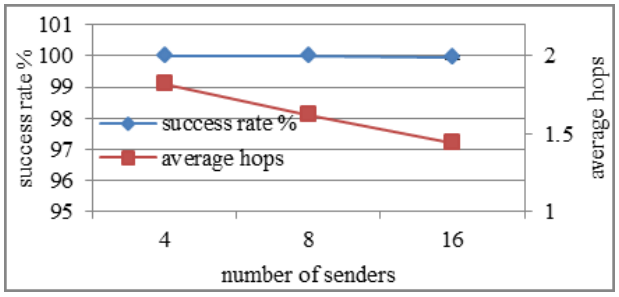


Figure 7A. Success Rate % and Avg Hops vs Senders

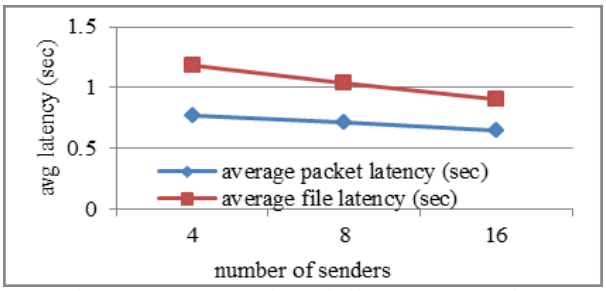


Figure 7B. Average Packet and File Latency vs Senders

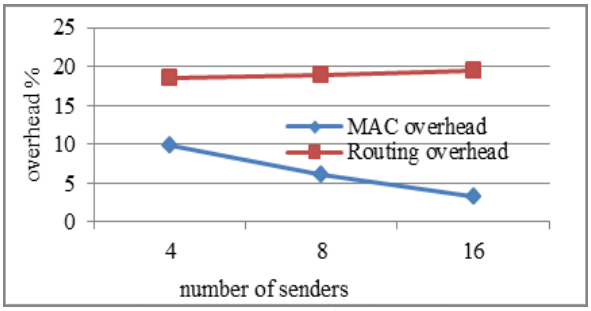


Figure 7C. Control Overhead vs Senders

The MAC overhead shows a decrease with increasing number of senders, because when there are fewer data packets to send (with less senders) the MAC still sends maintenance packets, thus the ratio of control bits to the

total bits that travelled the network, shows a decrease when there are more data packets in the network. The routing overhead records a very slight increase (around 1%) with increasing senders, which can be attributed to more route maintenance which will be triggered to correctly route the high amount traffic generated.

B. 50 Nodes Scenario

Figures 8A to 8C are the plots for the 50 UAV scenario with 10 clusters. The number of UAVs sending 1 MByte file simultaneously was varied from 10, 20 to 40. Thus in the case of the 40 senders, all CCs were sending 1 MByte files to the CHs simultaneously.

The success rate in graph A shows a slight drop to around 99.7 % as the senders increased, which shows the reliability in data transfer of the proposed solution and its scalability as the number of surveillance nodes and data sending nodes increased. The average hops which is plotted along with success rate graph does not show a linear

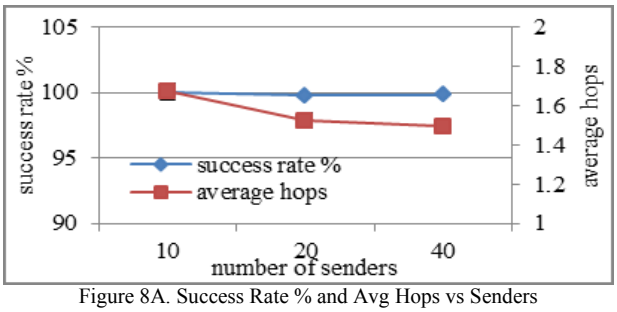


Figure 8A. Success Rate % and Avg Hops vs Senders

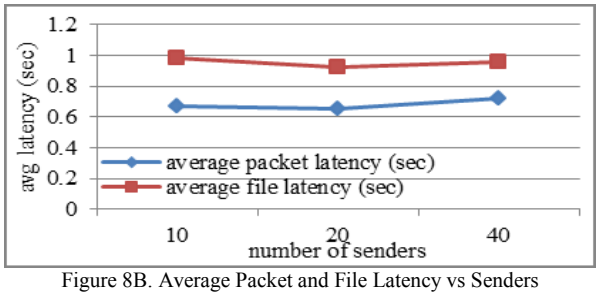


Figure 8B. Average Packet and File Latency vs Senders

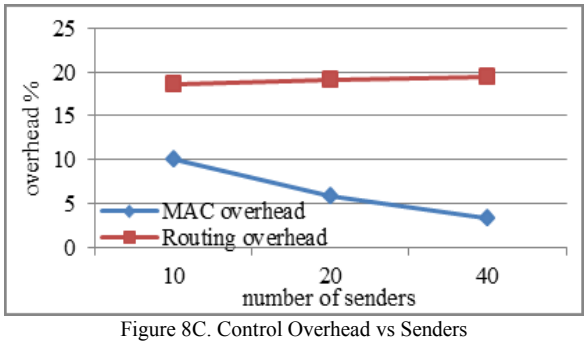


Figure 8C. Control Overhead vs Senders

decrease as in Fig 5 graph A. This is again attributed to the random selection in sending nodes. The first 10 senders were on an average of 1.7 hops from the CH, the added 10 senders for the 20 node case reduced the average hops to slightly above 1.5, and the last 20 senders brought the average hops to 1.5.

Figure 8B reflects the impact of the average hops in the packet and file delivery latency. There is drop when the senders increase from 10 to 20, this is because the average hops has a steep decrease from 1.7 to 1.5. However the average hops drops very slightly when senders are increased from 20 to 40 nodes, this and the fact that there is more traffic and more buffering by the nodes, the packet and file latency increase with increase in senders from 20 to 40.

The MAC and routing overhead in Figure 8C show a similar trend as observed in Figure 7. Though the number of nodes has increases, control traffic is calculated as a ratio of control traffic to total traffic in the network during the time that the files are being delivered.

C. 75 Nodes Scenario

Figures 9A, 9B and 9C are the performance plots for the test scenario with a total of 75 UAVs and 15 clusters, the number of sending nodes was varied from 15, 30 to 60. Hence again when 60 nodes are sending 1 Mbyte file it is the case of all CCs sending traffic to the CHs. The success rate dropped to around 98.7% with increasing number of senders – reflecting the robustness of the proposed solution and its scalability to increasing UAVs and increasing number of senders. The plot of the average hops again shows a decrease from 1.55 to 1.47 as the number of senders selected randomly to send the traffic to the CH was increased.

Figure 9B is the plot for the packet and file latency. The plot shows an increase because the change in the average hops was 0.06 as the number of senders was increased. The latency trends reflect the average hops trend. Figure 7C which is the plot of the MAC and routing overhead has a similar trend as noted for the 20 and 50 node scenarios.

Summarizing, the performance graphs indicate the high robustness of the proposed solutions to highly mobile and stressful MANET conditions. The continually high value of success rate despite the increase in the network size and the increase in the number of sending nodes indicate the reliability of the proposed solutions and its scalability. The packet and file latencies never exceeded 0.8 seconds and 1.2 seconds respectively in the three network setups. This indicates the robustness of the scheduling algorithm.

The overheads noted have similar trends and show very little difference as they were calculated as a ratio of the traffic in the network. The senders in each case were a quarter of the CCs, half of the CCS and the rest of the CCs. The control traffic increases with the increase in the number of nodes in a scenario, but as it is expressed as a ratio of all the traffic in the network including the data traffic, and due to the ratio of senders being consistent in all scenarios, this value can be noticed to be very close in all scenarios.

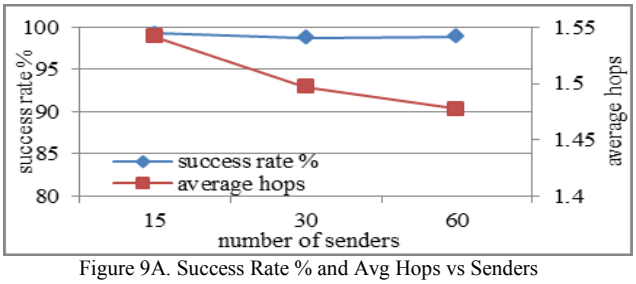


Figure 9A. Success Rate % and Avg Hops vs Senders

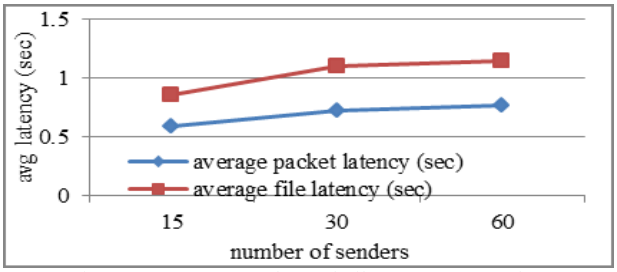


Figure 9B. Average Packet and File Latency vs Senders

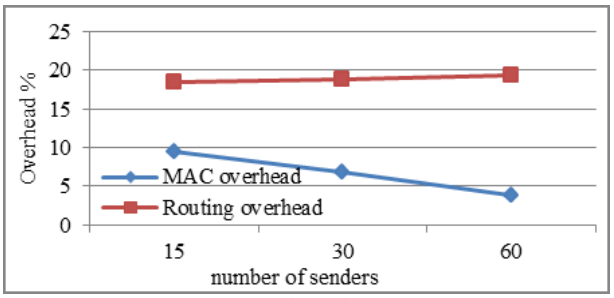


Figure 9C. Control Overhead vs Senders

VI. CONLUSION

Surveillance networks are critical tactical applications, and hence require special consideration during solution design. The primary goal in surveillance networks of UAVs is to collect the captured data reliably at few nodes, and with low latencies. In this work we presented a solution that uses an integrated approach where MAC, routing and scheduling are based off a single algorithm and use a single address. The design leads to a new MANET architecture that has performance advantages over traditional approaches and provides a low complexity yet robust and scalable solution.

The solution was evaluated in a UAV surveillance network of varying sizes of 20, 50 and 75 nodes. In each case the numbers of simultaneous 1 MByte file senders were increased from one quarter to one half to all of the remaining nodes besides the aggregation nodes. This was a highly stressful test case. The results achieved under such stress situations were very good. The drop in reliable and timely delivery was very low as the numbers of senders were increased. These results thus validate the use of the solution to such critical tactical applications.

The proposed solution has several tunable parameters as the MMT algorithm allows such capabilities. These capabilities are optimizing the cluster size, determining the number of VIDs to allow for nodes, decisions by nodes to join different clusters or have several branches under one cluster, length the tree branches and so on. The architecture has the feature to allow considering applications criteria and physical layer constraints while determining the paths. The information could be used for improved system design. This is due to the structure and positioning of the communications layer between the applications layer and physical layer. The solution is transparent to layer 3 and hence will not be impacted during IPv4 to Ipv6 transition or to any other layer 3 protocol. It can thus interwork with existing systems and their protocol structures.

ACKNOWLEDGMENT

This work was partly by funded by AFRL, Rome NY under contract no. 30822, and partly from ONR.

REFERENCES

[1] Bill Huba, Nirmala Shenoy, " Airborne Surveillance Networks with Directional Antennas", IARIA *International conference on Computers and network Systems, ICNS 2012*, St Maarten Islands 24-30 March 2012, Netherlands

[2] R. Nelson, L. Kleinrock, Spatial-TDMA: A collision-free multihop channel access protocol, IEEE Transactions on Communications 33 (1985) 934–944.

[3] I. Martinez and J. Altuna, Influence of directional antennas in STDMA ad hoc network schedule creation," in International Workshop on Wireless Ad-hoc Networks, (London, UK), 2005.

[4] S. G. Fernandez, On the performance of STDMA Link Scheduling and Switched Beamforming Antennas in Wireless Mesh Networks, Master's thesis, King's College London, London, United Kingdom, 2009.

[5] J. Grönkvist and A. Hansson, "Comparison between graph-based and interference-based STDMA scheduling," in *MobiHoc*, 2001.

[6] J. Grönkvist, Traffic controlled spatial reuse TDMA in multi-hop radio networks, in: The 9th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 1998, pp. 1203–1207.

[7] J. Grönkvist, Assignment methods for spatial reuse TDMA, in: First Annual Workshop on Mobile and Ad Hoc Networking and Computing, 2000, pp. 119–124.

[8] J. Grönkvist, A. Hansson, J. Nilsson, A comparison of access methods for multi-hop ad hoc radio networks, in: IEEE Vehicular Technology Conference, 2000, pp. 1435–1439.

[9] A. Dhamdhere, J. Grönkvist, "Joint Node and Link Assignment in an STDMA Network," In Proc. IEEE Vehicular Technology Conference, 22-25 April 2007, pp. 1066 – 1070.

[10] J. Grönkvist, "Novel Assignment Strategies for Spatial Reuse TDMA in Wireless Ad hoc Networks," Wireless Networks, Springer Netherlands, ISSN 1022-0038, vol. 12, no. 2, pp. 255 – 265, 2006.

[11] J. Grönkvist, Jan Nilsson, and D. Yuan, "Throughput of optimal spatial reuse TDMA for wireless ad-hoc networks," in Proc. of VTC 2004-Spring, Milan, Italy, May 2004.

[12] Gerla M. and Tsai J., "Multicluster, mobile, multimedia radio network," *Wirel. Netw.*, vol. 1, pp. 255-265, 1995

[13] Lian, J.; Agnew, G.B. and Naik, S., "A variable degree based clustering algorithm for networks," *Computer Communications and Networks, 2003. ICCCN 2003. Proceedings. The 12th International Conference on* , vol., no., pp. 465-470, 20-22 Oct. 2003

[14] Amis, A. D., Prakash, R., Vuong, T.H.P. and Huynh, D.T., "Max-min d-cluster formation in wireless ad hoc networks," INFOCOM 2000. vol.1, no., pp.32-41 vol.1.

[15] Lin, C.R.; Gerla, M., "Adaptive clustering for mobile wireless networks," *Selected Areas in Communications, IEEE Journal on* , vol.15, no.7, pp.1265-1275, Sep 1997

[16] Basagni, S., "Distributed and mobility-adaptive clustering for multimedia support in multi-hop wireless networks," *Vehicular Technology Conference, 1999. VTC 1999 - Fall. IEEE VTS 50th* , vol.2, no., pp.889-893 vol.2, 1999

[17] Hong X., Xu K., and Gerla M., "Scalable Routing Protocols for Mobile Ad Hoc Networks", *IEEE Network Journal,* July/Aug 2002, Vol 16, issue 4, pp 11-2.

[18] Abolhasan M., Wysocki T. and Dutkiewicz E., "A review of routing protocols for mobile ad hoc networks", *Journal of ad hoc networks,* Elsevier publications, 2004

[19] Qin L. Kunz T., "Survey on Mobile Ad Hoc Network Routing Protocols and Cross-Layer Design" Technical Report Systems and Computer Engineering, Carleton University, August 2004

[20] Abolhasan M., Wysocki T., Dutkiewicz E., "A review of routing protocols for mobile ad hoc networks", Journal of ad hoc networks, Elsevier publications, 2004

[21] Daniel L., "A comprehensive overview about selected Ad hoc networking routing protocols", Technical Report, Department of Computer Science, Technische Universitat, Munchen, Germany

[22] Royer E. M., C.-K. Toh, "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks", IEEE Personal Communications Magazine, April 1999, pages 46-55.

[23] Perkins C., E., E. M. Royer, and S. R. Das, "Ad Hoc On-Demand Distance Vector (AODV) Routing,. IETF Mobile Ad Hoc Networks Working Group", IETF RFC 3561

[24] Johnson D. B., D. A. Maltz, and Y-C Hu., "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)," IETF Mobile Ad Hoc Networks Working Group, Internet Draft, 24 February 2003

[25] Clausen T., Ed., P. Jacquet, .Optimized Link State Routing Protocol (OLSR)., Network Working Group, Request for Comments: 3626

[26] Das S., R. Castaneda, and J. Yan, "Simulation-Based Performance Evaluation of Routing Protocols for Mobile Ad Hoc Networks," Mobile Networks and Applications, 2000, Vol. 5, No. 3, pages 179-189

[27] Hong X., Kaixin Xu, Mario Gerla, "Scalable Routing Protocols for Mobile Ad Hoc Networks", IEEE Network Journal, July/Aug 2002, Vol 16, issue 4, pages 11-2.

[28] Pei G., M. Gerla, and T.-W. Chen, "Fisheye State Routing: A Routing Scheme for Ad Hoc Wireless Networks," IEEE International Conference on Communications, 2000, Vol 1, pages 70-74.

[29] Bellur B. and R. G. Ogier, "A Reliable, Efficient Topology Broadcast Protocol for Dynamic Networks," in Proc. IEEE INFOCOM '99, New York, March 1999.

[30] Santivanez C., R. Ramanathan, I. Stavrakakis, "Making Link-State Routing Scale for Ad Hoc Networks," in Proceedings of The 2001 ACM International Symposium on Mobile Ad Hoc Networking and Computing (Mobihoc2001), Long Beach, California, Oct. 2001.

[31] Das S.R., C.E. Perkins and E. M. Royer, "Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks," in Proceedings of IEEE INFOCOM 2000, Tel Aviv, Israel, Mar. 2000.

[32] Pei G., M. Gerla, X. Hong, and C. -C. Chiang, "A Wireless Hierarchical Routing Protocol with Group Mobility," in Proceedings of IEEE WCNC'99, New Orleans, LA, Sept. 1999.

[33] Haas Z.J. and M.R. Pearlman, "The Performance of Query Control Schemes for the Zone Routing Protocol," ACM/IEEE Transactions on Networking, vol. 9, no. 4, August 2001, pp. 427-438.

[34] Pei G., M. Gerla and X. Hong, "LANMAR: Landmark Routing for Large Scale Wireless Ad Hoc Networks with Group Mobility," in Proceedings of IEEE/ACM MobiHOC 2000, Boston, MA, Aug. 2000, pp. 11-18.

[35] Xu K., Hong, X., Gerla M., "An Ad hoc Network with Mobile Backbone", Communications, 2002. ICC 2002. IEEE International Conference on Communications, Volume 5, 28 April-2 May 2002 Page(s):3138 - 3143 vol.5

[36] Ramasubramanian V., Haas Z. J., Emin G¨un Sirer, "SHARP: A Hybrid Adaptive Routing Protocol for Mobile Ad Hoc Networks", Proceeding of Mobihoc 03, of the 4th ACM International Symposium on Mobile Ad Hoc Networking and Computing, Pages 303-314.

[37] Orakwue C., Al-Mousa Y., Martin N. and Shenoy N., "Cluster Based Time Division Multiple Access Scheme for Surveillance Networks using Directional Antennas" *ICSPCS, Brisbane Australia* Dec 2010.

[38] Huba W., Martin N., Al-Mousa Y., Orakwue C. and Shenoy N., "A Distributed Scheduler for Airborne Backbone Networks with Directional Antennas", *ComsNets,* Bangalore India 2011

[39] Martin N., Al-Mousa Y. and Shenoy N., "An Integrated Routing and Medium Access Control Framework for Surveillance Networks of Mobile Devices" *12th ICDCN 2011*, Springer Verlag, pp 315-323.

[40] Shenoy N., Yin Pan, Darren Narayan, David Ross, Carl Lutzer, "Route Robustness of a Multi-meshed Tree Routing Scheme for Internet MANETs", Proceeding of IEEE Globecom 2005. 28 Nov – 2nd Dec. 2005 St Louis.

[41] Pudlewski S., Shenoy N., Al Mousa Y., "A Hybrid Multi Meshed Tree Routing protocol for wireless ad hoc networks", Second IEEE International Workshop on *Enabling Technologies and Standards for Wireless Mesh Networking,* September 29, 2008. Atlanta, GA, USA

A Performability Modeling Framework Considering Service Components Deployment

Razib Hayat Khan
Department of Telematics
NTNU, Norway
rkhan@item.ntnu.no

Fumio Machida
Service Platform Research
NEC, Japan
h-machida@ab.jp.nec.com

Poul E. Heegaard
Department of Telematics
NTNU, Norway
poul.heegaard@item.ntnu.no

Kishor S. Trivedi
Department of ECE
Duke University, NC, USA
kst@ee.duke.edu

***Abstract*- The analysis of the system behavior from the pure performance viewpoint tends to be optimistic since it ignores failure and repair behavior of the system components. On the other hand, pure dependability analysis tends to be too conservative since performance considerations are not taken into account. The ideal way is to conduct the modeling of performance and dependability behavior of the distributed system jointly for assessing the anticipated system performance in the presence of system components failure and recovery. However, design and evaluation of the combined model of a distributed system for performance and dependability analysis is burdensome and challenging. Focusing on the above contemplation, we introduce a framework to provide tool based support for performability modeling of a distributed software system that proposes an automated transformation process from the high level Unified Modeling Language (UML) notation to the Stochastic Reward Net (SRN) model and solves the model for early assessment of a software performability parameters. UML provides enhanced architectural modeling capabilities but it is not a formal language and does not convey formal semantics or syntax. We present the precise semantics of UML models by formalizing the concept in the temporal logic compositional temporal logic of actions (cTLA). cTLA describes various forms of actions through an assortment of operators and techniques which fit excellently with UML models applied in this work and also provides the support for incremental model checking. The applicability of our framework is demonstrated in the context of performability modeling of a distributed system to show the deviation in the system performance against the failure of system components.**

Keywords: UML; SRN; Performability; Deployment; Reusability

I. INTRODUCTION

Conducting performance modeling of a distributed system separately from the dependability modeling fails to asses the anticipated system performance in the presence of system components failure and recovery. System dynamics is affected by any state changes of the system components due to failure and recovery. This introduces the concept of performability that considers the behavioral change of the system components due to failures and also reveals how this behavioral change affects the system performance. But to design a composite model for a distributed system, perfect modeling of the overall system behavior is essential and sometimes very unwieldy. A distributed system behavior is normally realized by the several objects that are physically disseminated. The overall system behavior is maintained by the partial behavior of the distributed objects of the system [14]. So it is essential to model the distributed objects behavior perfectly for appropriate demonstration of the system dynamics and to conduct the performability evaluation [14]. Hence, we adopt UML collaboration, state machine, deployment, and activity oriented approach as UML is the most commonly used specification language which models both the system requirements and qualitative behavior through an assortment of notations [5] [14]. The way we utilize the UML collaboration and activity diagram to capture the system dynamics, provides the opportunity to reuse the software components. The specifications of collaboration are given as coherent, self-contained building blocks [14]. Reusability of the software component is achieved by designing the collaborative building block which is used as main specification unit in this work. Collaboration with help of activity diagram illustrates the complete behavior of a software system which includes both the local behavior among the participants and necessary interactions among them. Moreover, for specifying deployment mapping of service components, the performability modeling framework considers system execution architecture through UML deployment diagram. Considering system execution architecture while designing the framework resolves the bottleneck of the deployment mapping of service components by revealing a better allocation of service components to the physical nodes [13]. This requires an efficient approach to deploy the service components on the available hosts of a distributed environment to achieve preferably high performance and low cost levels [14]. Later on, UML State machine (STM) diagram is employed in this framework to capture system components behavior with respect to failure and repair events.

In order to guarantee the precise understanding and correctness of the model, the approach requires formal reasoning on the semantics of the language used and to maintain the consistency of the models specification. Temporal logic is a suitable option for that. In particular, the properties of super position supported by cTLA [19] make it possible to describe systems from different view points by individual processes that are superimposed. In this work, we focus on the cTLA that allows us formalizing the collaborative service specifications given by UML activities and also to define the formal semantics of the UML

deployment diagram and STM model precisely. By expressing collaborations as cTLA processes, we can ensure that a composed service maintains the properties of the individual collaborations it is composed of. The semantic definition of collaboration, activity, deployment, and STM model in the form of temporal logic is implemented as a transformation tool [20] which produces TLA^+ modules. These modules may then be used as input for the model checker TLC for syntactic analysis [20].

Furthermore, UML models are annotated according to the *UML profile for MARTE* [7] and *UML profile for Modeling Quality of Service and Fault Tolerance Characteristics* [13] to include quantitative system parameters necessary for performability evaluation. UML specification styles are applied to generate the SRN model automatically following the model transformation rules where model synchronization between the performance and dependability SRN model is achieved by defining guard functions (a special property of the SRN model [6]). This synchronization thus helps to properly model the system performance with respect to any state changes in the system due to components failure [1] [2].

Over decades several performability modeling techniques have been considered such as Markov models, SPN (Stochastic Petri Nets) and SRN [4]. Among all of these, we will focus on the SRN as performability model generated by our framework due to its prominent and interesting properties such as priorities assignment in transitions, presence of guard functions for enabling transitions that can use entire state of the net rather than a particular state, marking dependent arc multiplicity that can change the structure of the net, marking dependent firing rates, and reward rates defined at the net level [6].

Several approaches have been pursued to accomplish a performability analysis model from a system design specification. Sato et al. develop a set of Markov models, for computing the performance and the reliability of Web services and detecting bottlenecks [9]. Another initiative focuses on model-based analysis of performability of mobile software systems by proposing a general methodology that starts from design artifacts expressed in a UML-based notation. Inferred performability models are formed based on the Stochastic Activity Networks notation [10]. Subsequent effort proposes a methodology for the modeling, verification, and performance evaluation of communication components of a distributed application building software which translates UML 2.0 specifications into executable simulation models [11]. Gonczy et al. mentioned a method for high-level UML models of service configurations captured by a UML profile dedicated to service design; performability models are derived by automated model transformations for the PEPA toolkit in order to assess the cost of fault tolerance techniques in terms of performance [12]. However, most of the existing approaches do not consider the fact of how to conduct the system modeling to delineate system functional behavior while generating the performability model using reusable software components. The framework introduced in this work is superior to the existing approaches that have been realized by UML specification style as reusable building block to characterize a system dynamics. The purpose of the reusable building block is twofold: to express the local behavior of several components and to capture the interaction between them. This provides the excellent opportunity to reuse the building blocks, as the interaction among the several components can be encapsulated within one self-contained building block [14]. This reusability provides the means to design a new system's behavior rapidly utilizing the existing building blocks according to the specification. This helps to start the development process from scratch which in turn facilitates the swelling of productivity and quality in accordance with the reduction in time and cost [2]. Moreover, the ensuing deployment mapping given by our framework has greater impact to satisfy QoS requirements provided by the system. The target in this work is to deal with vector of QoS instead of confining them in one dimension. Our provided deployment logic is definitely capable of handling any properties of the service as long as a cost function for the specific property can be produced. The defined cost function is able to react in accordance with the changing size of search space of available hosts presented in the execution environment to assure an efficient deployment mapping [14]. In addition, the separation of performance and dependability modeling view and the introduction of model synchronization to synchronize the two views activities using guard functions relinquishes the complex and unwieldy affect in performability modeling and evaluation of large and multifaceted systems [1].

The objective of this paper is to provide a tool based support for the performability modeling of a distributed system to allow modeling of the performance and dependability related behavior in a combined and automated way. This in turn allows not only to model functional attributes of the service provided by the system but also to investigate dependability attributes to reflect how the changes in the dependability attributes affect the system overall performance. For ease of understanding the complexity behind the modeling of performability attributes, our modeling framework works in two different views such as performance modeling view and dependability modeling view. The framework achieves its objective by maintaining harmonization between performance and dependability modeling view with the support of model synchronization. The paper is organized as follows: Section II introduces our performability modeling framework, Section III depicts UML model description, Section IV describes formalization of UML models, Section V explains service components deployment issue, Section VI clarifies UML models annotations, Section VII delineates model transformation rules, Section VIII introduces the model synchronization mechanism, Section IX describes the hierarchical method for mean time to failure (MTTF) calculation, Section X indicates the tool based support of the modeling framework, Section XI illustrates the case study, and Section XII delineates the concluding remarks with future directions.

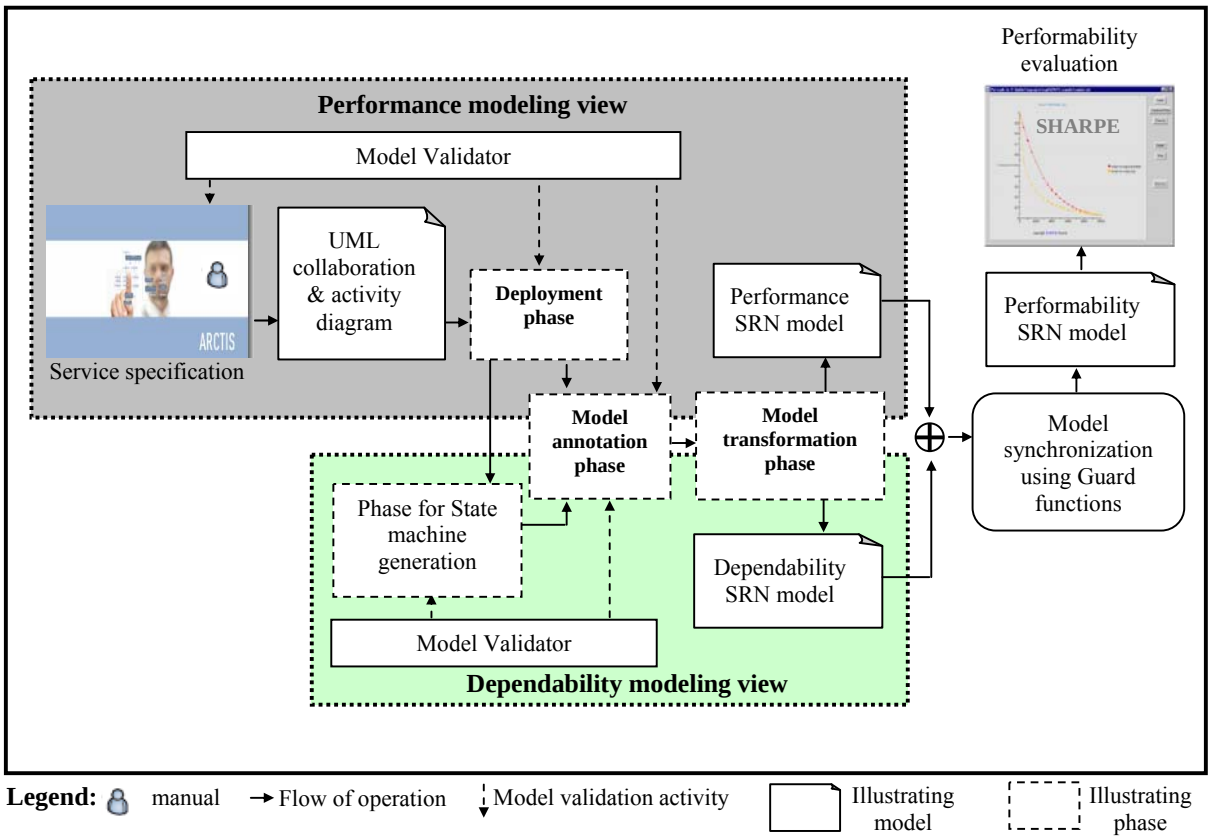


Figure 1. Proposed performability modeling framework

II. OVERVIEW OF PROPOSED FRAMEWORK

Our performability framework is composed of 2 views: performance modeling view and dependability modeling view. The performance modeling view mainly focuses on capturing the system's dynamics to deliver certain services deployed on a distributed system. The performance modeling view is divided into 4 steps shown in Fig. 1 where the service specification step is the part of Arctis tool suite which is integrated as plug-ins into the eclipse IDE [15]. Arctis focuses on the abstract, reusable service specifications that are composed of UML 2.2 collaborations and activities [15]. It uses collaborative building blocks to create comprehensive services through composition. In order to support the construction of building block consisting of collaborations and activities, Arctis offers special actions and wizards.

In the first step of performance modeling view, a developer consults a library to check if an already existing basic building block or collaboration between several blocks solves a certain task. Missing blocks can also be created from existing building blocks and stored in the library for later reuse. The building blocks are expressed as UML models. The structural aspect, for example the service components and their multiplicity, is expressed by means of UML 2.2 collaborations. For the detailed internal behavior, UML 2.2 activities have been used. The building blocks are combined into more comprehensive service by composition to specify the detailed behavior of how the different events of collaborations are composed. For this composition, UML collaborations and activities are used complementary to each other [15]. In the deployment phase, the deployment diagram of our proposed system is delineated and the relationship between system components and collaborations is outlined to describe how the service is delivered by the joint behavior of the system components. In the model annotation phase, performance information is incorporated into the UML activity diagram and deployment diagram according to the *UML profile for MARTE* [8]. The model transformation phase is devoted to automate generation of a SRN model following the model transformation rules. The SRN model generated in this view is called performance SRN.

The dependability modeling view is responsible for capturing any state changes in the system because of failure and recovery behavior of system components. The dependability modeling view is composed of three steps shown in Fig. 1. In the first step, UML STM diagram is used to describe the state transitions of software and hardware components of the system to capture the failure and recovery events. In the model annotation phase, dependability parameters are incorporated into the STM diagram according to *UML profile for Modeling Quality of Service and Fault Tolerance Characteristics & Mechanisms Specification* [13]. The model transformation phase reflects the automated generation of the SRN model from the STM

diagram following the model transformation rules. The SRN model generated in this view is called dependability SRN.

The model synchronization is used as glue between performance SRN and dependability SRN. The synchronization task guides the performance SRN model to synchronize with the dependability SRN model by identifying the transitions in the dependability SRN. The synchronization between performance and dependability SRN is achieved by defining the guard functions. Once the performance SRN model is synchronized with dependability SRN model, a merged SRN model will be obtained and various performability measures can be evaluated from the merged model using the software package SHARPE [16].

III. UML BASED SYSTEM DESCRIPTION

A. Construction of collaborative building blocks

The performability modeling framework utilizes collaboration as main entity. Collaboration is an illustration of the relationship and interaction among software objects in the UML. Objects are shown as rectangles with naming label inside. The relationships between the objects are shown in a oval connecting the rectangles [5]. The specifications for collaborations are given as coherent, self-contained reusable building blocks. The structure of the building block is described by UML 2.2 collaboration. The building block declares the participants (as collaboration roles) and connection between them. The internal behavior of building block is described by the UML activity. It is declared as the classifier behavior of the collaboration and has one activity partition for each collaboration role in the structural description. For each collaboration, the activity declares a corresponding call behavior action referring to the activities of the employed building blocks. For example, the general structure of the building block *t* is given in Fig. 2 where it only declares the participants *A* and *B* as collaboration roles and the connection between them is defined as collaboration t_x *(*$x=1 \ldots n_{AB}$ *(*number of collaborations between collaboration roles *A & B))*. The internal behavior of the same building block is shown in Fig. 3(b). The activity $transfer_{ij}$ (where *ij = AB*) describes the behavior of the corresponding collaboration. It has one activity partition for each collaboration role: *A* and *B*. Activities base their semantics on token flow [2]. The activity starts by forwarding a token when there is a response (indicated by the streaming pin *res*) to transfer

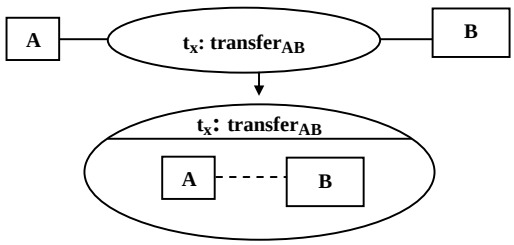


Figure 2. Structure of the Building block

from participant *A* to *B*. The token is then transferred by the participant *A* to participant *B* (represented by the call operation action *forward*) after completion of the processing by the collaboration role *A*. After getting the response of the participant *A,* the participant *B* starts the processing of the request (indicated by the streaming pin *req*).

In order to generate the performability model, the structural information about how the collaborations are composed is not sufficient. It is necessary to specify the detailed behavior of how the different events of collaborations are composed so that the desired overall system behavior can be obtained. For the composition, UML collaborations and activities are used complementary to each other. UML collaborations focus on the role binding and structural aspect, while UML activities complement this by covering also the behavioral aspect for composition. Therefore, the activity contains a separate call behavior action for all collaborations of the system. Collaboration is represented by connecting their input and output pins. Arbitrary logic between pins may be used to synchronize the building block events and transfer data between them. By connecting the individual input and output pins of the call behavior actions, the events occurring in different collaborations can be coupled with each other. Semantics of the different kinds of pins are given in more details in [14]. For example, the detail behavior and composition of the collaboration is given in following Fig. 3(a). The initial node (●) indicates the starting of the activity. The activity is started from the participant *A*. After being activated, each participant starts its processing of request which is mentioned by call operation action Pr_i *(*$Processing_i$*, where i = A, B & C)*. Completion of the processing by the participants are mentioned by the call operation action Prd_i *(*$Processing\_done_i$*, where i = A, B & C).* After completion of the processing, the response is delivered to the corresponding participant. When the processing of the task by the participant *A* completes, the response (indicated by streaming pin *res)* is transferred to the participant *B* mentioned by collaboration *t:* $transfer_{ij}$ (*where ij = AB)* and participant *B* starts the processing of the request (indicated by streaming pin *req).* After completion of the processing, participant *B* transfers the response to the participant *C* mentioned by collaboration *t:* $transfer_{ij}$ (*where ij = BC)*. Participant *C* starts the processing after receiving the response from *B* and activity is terminated after completion of the processing which is illustrated by the terminating node (◉).

B. Modeling failure & repair behavior of software & hardware component using UML STM

State transitions of a system element are described using UML STM diagram. In an STM, a state is depicted as a rectangle and a transition from one state to another is represented by an arrow [5]. In this work, STM is used to describe the failure and recovery behavior of software and hardware components.

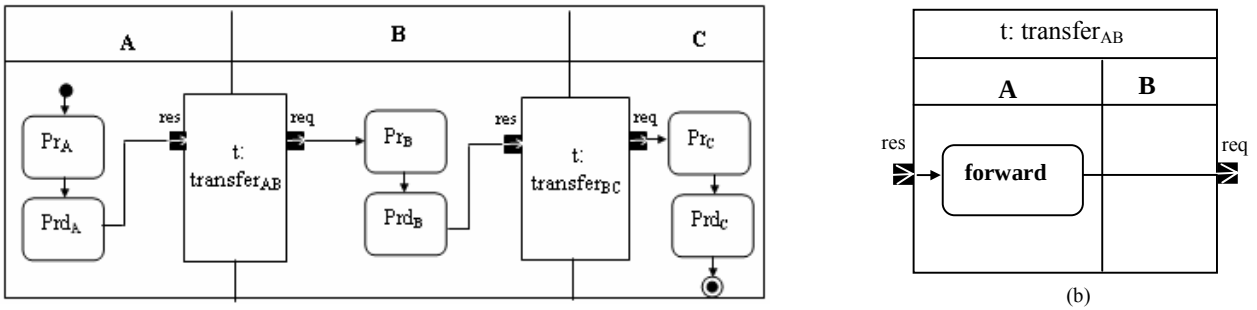


Figure 3. (a) Detail behavior of the event of the collaboration using activity (b) internal behavior of the collaboration

The STM of software process is shown in Fig. 4(a). The initial node (●) indicates the starting of the operation of software process. Then the process enters **Running** state. **Running** is the only available state in the STM. If the software process fails during the operation, the process enters **Failed** state. When the failure is detected by the external monitoring service the software process enters **Recovery** state and the repair operation will be started. When the failure of the process is recovered the software process returns to **Running** state.

The STM of hardware component is shown in Fig. 4(b). The initial node (●) indicates the starting of the operation of hardware component. Then the component enters **Running** state. **Running** is the only available state here. If the active component fails during the operation and the hot standby component is available, the standby component will take charge and the component operation will be continued. When any failure (whether active component or standby component) incurs, the recovery operation will be performed.

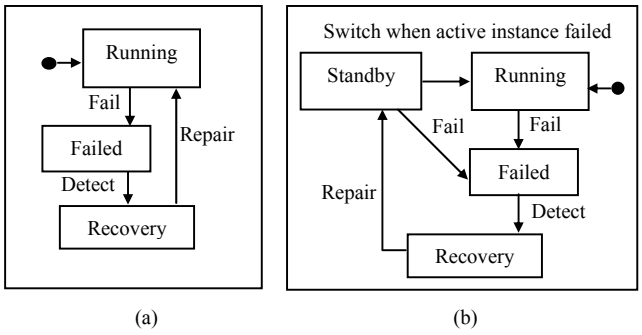


Figure 4. (a) STM of Software process (b) STM of Hardware component

IV. FORMALIZING UML DIAGRAM

So far we introduced the UML diagrams in a descriptive and informal way. In order to understand the precise formalism of the UML models and for the correct way of model transformation, we need to present the UML models with the help of formal semantics. The formal semantics of UML models thus help us implementing the models very efficiently for providing the tool based support of our framework. Before introducing the formalization of the UML models, at first, we illustrate the temporal logic, more specifically compositional Temporal Logic of Actions (cTLA) that will be applied to formalize the UML models. We illustrate in this paper the formal representation of the state machine model. Formalization of other UML models such as collaboration, activity, and deployment diagram and the alignment between UML models and cTLA (which is beyond the scope of this paper) have already been mentioned in [22].

A. *Compositional Temporal Logic of Action (cTLA)*

Lamport's Temporal Logic of Actions (TLA, [21]) is a linear-time temporal logic modeling the system behavior where the system behavior is realized by a set of considerably large number of state sequences [s_0, s_1, s_2 . .] [23]. Thus, the TLA formalisms are applied nicely to define the state machines formally produced by our framework which, in the end, also models considerably long sequences of states s_i starting with an initial state s_0. Compositional TLA (cTLA, [22]) was originated from TLA to offer more easily comprehensible formalisms and proposes a more supple composition of specifications. The concept of process is basically introduced by a cTLA. A cTLA process describes system behavior as the notion of state transition systems [23].

B. *Formalizing state machine diagram using cTLA*

We sketch the cTLA model of STM in Fig. 5 by the specification of software process dependability behavior illustrated in Fig. 4(a) [23]. The header *Software* declares the name of the process type. *Events* is an expression defined as constant record type. The state space is modeled by a set of variables like *state* or *Queue*. Predicate *INIT* specifies the subset of initial states. The state transition systems are mentioned by actions (*e.g., enqueue, dequeu*) which are realized as pairs of current and next states describing a set of transitions each. The current sate is defined as a variable in simple form (*e.g., state*), while the next state is mentioned by the prime form (*e.g., state´*). Variables which won't be changed by an action are listed by the statement UNCHANGED [23]. State transition system is defined by the body of a cTLA process type. One cTLA process represents one state machine that mentions a set of TLA state sequences. The first state s_0 of each modeled state

sequence has to fulfil the initial condition *INIT*. The state changes [s_i, s_{i+1}] either correspond with a process action or with a so-called stuttering step in which the current and the next states are equal (i.e., $s_i = s_{i+1}$) [23]. Incoming events are

```
PROCESS Software ()

CONSTANTS
        Events ≜ {Fail, Detect, Recovery};

VARIABLES
        state: {initState, running, failed, recovered};
        Queue: QUEUE of Events;

INIT ≜ state = initState ⋀ Queue = EMPTY;

ACTIONS
    enqueue (addEvent: Events) ≜
    Queue´ = Queue ◦ addEvent ⋀ state ≠ initState ⋀
                UNCHANGED〈state〉
    dequeue (fetchEvent: Events) ≜
    Queue ≠ EMPTY ⋀ fetchEvent = FIRST(Queue) ⋀
                UNCHANGED 〈state〉
        Initial ≜ state = initState ⋀ state´= idle ⋀
                UNCHANGED 〈Queue〉
        stateseq ≜ state = idle ⋀ state´ = running ⋀
                UNCHANGED 〈Queue〉
 …….. state change for other actions ……

WF: dequeue, initial;
```

Figure 5. cTLA process of Software component

inserted into the data structure *addEvent*, which is a sequence of events. The operator ◦ denotes the concatenation of queue elements. Events are added to the queue by the action *enqueue*, which takes incoming events as action parameters [23]. Retrieving events are modeled by the data structure *fetchEvent* where the first element is obtained by the operations FIRST(). Events are retrieved from the queue by the action *dequeue* which takes retrieving events as action parameters. An initial transition initiates from an initial pseudo state (*initState*) and its execution is associated with the starting of the state machine. Exactly one initial transition is linked with each state machine [23]. A cTLA variable *state* describes the control state by expressing them through the control state identifiers. *Stateseq* captures the current and next state and starts from initial state of the STM diagram. In order to conduct an action in a lively manner, we can associate actions with weak and strong fairness properties. In particular, weak fairness forces the execution of an activity as if it were enabled continuously. Strong fairness forces the execution even if the action is sometimes disabled [23]. The last statement WF: dequeue, initial,... lists the actions that have to be carried out in a way which ensures week fairness property [23].

V. DEPLOYMENT DIAGRAM & STATING RELATION BETWEEN SYSTEM & SERVICE COMPONENT

We model the system as collection of *N* interconnected physical nodes. Our objective is to find a deployment mapping for this execution environment for a set of service components available for deployment that comprises the service. Deployment mapping M can be defined as [M=($C \rightarrow N$)] between a number of service components instances *C*, onto physical nodes *N*. We consider three types of requirements in the deployment problem where the term cost is introduced to capture several non-functional requirements; those are later on, utilized to conduct performance evaluation of the systems: (1) Service components have execution costs, (2) Collaborations have communication costs and costs for running of background process known as overhead cost, (3) Some of the service components can be restricted in the deployment mapping to specific physical nodes which are called bound components.

Furthermore, we consider identical physical nodes that are interconnected in a full-mesh and are capable of hosting service components with unlimited processing demand. We observe the processing cost that physical nodes impose while hosting the service components and also the target balancing of cost among the physical nodes available in the network. Communication costs are considered if collaboration between two service components happens remotely, i.e. it happens between two physical nodes [18]. In other words, if two service components are placed onto the same physical node the communication cost between them will be ignored. This holds for the case study that is conducted in this paper. This is not generally true, and it is not a limiting factor of our framework. The cost for executing the background process for conducting the communication between the collaboration roles is always considerable no matter whether the collaboration roles deploy on the same or different physical nodes. Using the above specified input, the deployment logic provides an optimal deployment architecture taking into account the QoS requirements for the service components providing the specified services. We then define the objective of the deployment logic as obtaining an efficient (low-cost, if possible optimum) mapping of service components onto the physical nodes that satisfies the requirements in a reasonable time. The deployment mapping providing optimal deployment architecture is mentioned by the cost function F(M), that is a function that expresses the utility of deployment mapping of service components on the physical resources with their constraints and capabilities by satisfying non-functional requirements of the system. The cost function is designed to reflect the goal of balancing the execution cost and minimizing the communication cost. This is in turn utilized to achieve reduced task turnaround time by maximizing the utilization of system resources while minimizing any communication between processing

nodes. That will offer a high system throughput, taking into account the expected execution and inter-node communication requirements of the service components on the given hardware architecture [14]. The evaluation of cost function F(M) is mainly influenced by our way of service definition. A service is defined in our approach as a collaboration of total *E* service components labeled as c_i (where i = 1…. *E*) to be deployed and total *K* collaborations between them labeled as k_j, (where j = 1 … *K*). The execution cost of each service component can be labeled as f_{c_i} , the communication cost between the service components is labeled as f_{k_j} and the cost for executing the background process for conducting the communication between the service components is labeled as f_{B_j} . Accordingly, we will strive for an optimal solution of equally distributed cost among the processing nodes and the lowest cost possible, while taking into account the execution cost f_{c_i} , *i* = 1….*E*, communication cost f_{k_j} , *j* = 1….*K*, and cost for executing the background process f_{B_j} , *j* = 1….*K*. f_{c_i} , f_{k_j} , and f_{B_j} are derived from the service specification, thus the offered execution cost can be calculated as $\sum_{i=1}^{|E|} f_{c_i}$. This way, the logic can be aware of the target average cost *T* per physical node (*X*= total number of physical nodes) [18]:

$$T = \frac{1}{|X|}\sum_{i=1}^{|E|} f_{c_i} \qquad (1)$$

In order to cater for the communication cost f_{k_j} , of the collaboration k_j in the service, the function $q_0(M,c)$ is defined first [20]:

$$q_0(M,c) = \{n \in N \mid \exists(c \to n) \in M\} \qquad (2)$$

This means that $q_0(M,c)$ returns the physical node *n* from a vector of physical nodes *N* available in the network that host component in the list mapping M. Let collaboration $k_j = (c_1, c_2)$. The assumption in this paper is that, the communication cost of k_j is 0 (in general, it can be non-zero) if components c_1 and c_2 are collocated, i.e. $q_0(M,c_1) = q_0(M,c_2)$ and the cost is f_{k_j} if service components are otherwise (i.e., the collaboration is remote). Using an indicator function *I(x)*, which is 1 if *x* is true and 0 otherwise, this is expressed as $I(q_0(M,c_1) \neq q_0(M,c_2)) = 1$, if the collaboration is remote and 0 otherwise. In order to determine which collaboration k_j is remote, the set of mapping M is used. Given the indicator function, the overall communication cost of service, F_K(M), is the sum [20]:

$$F_K(M) = \sum_{j=1}^{|K|} I(q_0(M,k_{j,1}) \neq q_0(M,k_{j,2})) \cdot f_{k_j} \qquad (3)$$

Given a mapping M = {m_n} (where m_n is the set of service components at physical node *n*) the total load can be obtained as $\hat{l}_n = \sum_{c_i \in m_n} f_{c_i}$. Furthermore, the overall cost function F(M) becomes [20] (where I_j = 1, if k_j external or 0 if k_j internal to a node):

$$F(M) = \sum_{n=1}^{|X|} |\hat{l}_n - T| + F_K(M) + \sum_{j=1}^{|K|} f_{B_j} \qquad (4)$$

The absolute value $|\hat{l}_n - T|$ is used to penalize the deviation from the desired average load per node.

VI. ANNOTATION

In order to annotate the UML diagrams, the stereotype *saStep, computingResource, scheduler, QoSDimension*, and the tagged value *execTime, deadline, mean-time-to-repair, mean-time-between-failures,* and *schedPolicy* are used according to the *UML profile for MARTE* and *UML Profile for Modeling Quality of Service & Fault Tolerance Characteristics* [8] [13]. The stereotypes are the following:

- *saStep* defines a step that begins and ends when decisions about the allocation of system resources are made.
- *computingResource* represents either virtual or physical processing devices capable of storing and executing program code. Hence, its fundamental service is to compute.
- *scheduler* is a stereotype that brings access to a resource following a certain scheduling policy mentioned by tagged value *schedPolicy.*
- *QoSDimension* provides support for the quantification of QoS characteristics and attributes *mean-time-to-repair* and *mean-time-between-failures* [13].

The tagged values are the following:

- *execTime*: The duration of the execution time is mentioned by the tagged value *execTime* which is the average time in our case.
- *deadline* defines the maximum time bound on the completion of the particular execution segment that must be met.
- *mean-time-between-failures* defines the mean time of occurring a software and hardware instance failure
- *mean-time-to-repair* defines the mean time that is required to repair a software or hardware instance failure

We also introduce a new stereotype <<*transition*>> and three tag values *mean-time-to-stop, mean-time-to-start*, and *mean-time-to-failure-detect*.

- <<*transition*>> induces a state transition of a scenario.
- *mean-time-to-stop* defines the mean time that is required by a hardware instance to stop working
- *mean-time-to-start* states the mean time that is required by a hardware instance to start working
- *mean- time-to-failure-detect* defines the mean time that is required to detect failures in the system.

Fig. 6 illustrates an example annotated UML model using the activity diagram where the flow between P_A and d_A is annotated using stereotype *saStep* and tagged value

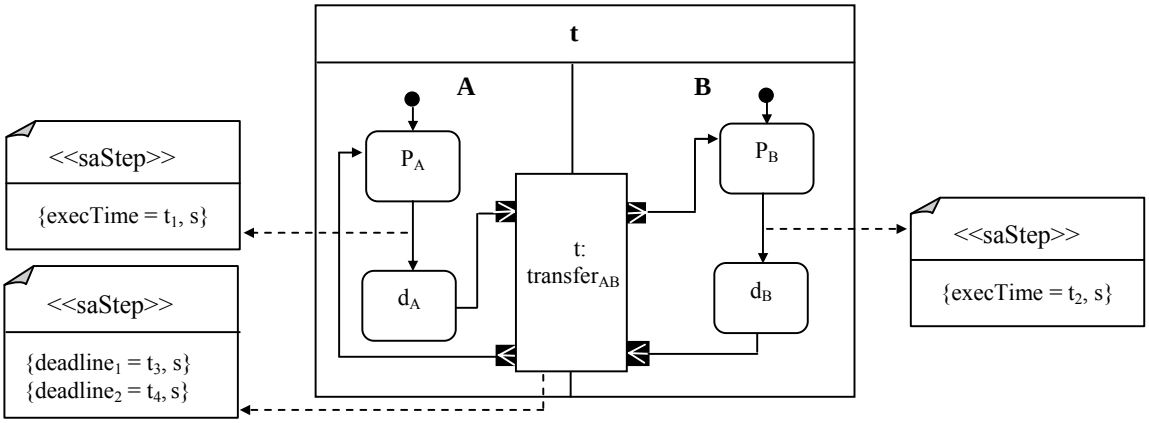


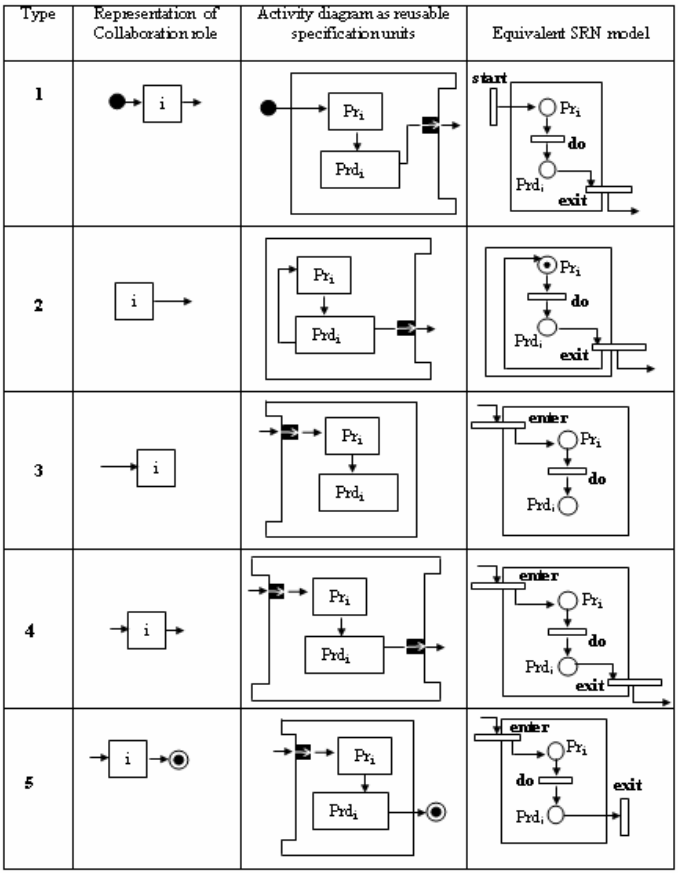
Figure 6. Annotated UML model

execTime which defines that after being deployed in an execution environment the collaboration role *A* needs t_1 seconds and collaboration role *B* needs t_2 seconds to complete their processing by the physical node. After completing the processing, communication between *A* and *B* is achieved in t_3 sec while the overhead time to conduct this communication is t_4 sec which is annotated using stereotype *saStep* and two instances of *deadline – deadline*$_1$ defines the communication time and *deadline*$_2$ is for overhead time.

VII. MODEL TRANSLATION

This section highlights the rules for the model translation from various UML models into SRN models. Since all the models will be translated into the SRN model, we will give a brief introduction about SRN model. SRN is based on the Generalized Stochastic Petri Net (GSPN) [4] and extends them further by introducing prominent extensions such as

TABLE I. SPECIFICATION OF REUSABLE UNITES AND EQUIVALENT SRN MODEL



guard function, reward function, and marking dependent firing rate [6]. A guard function is assigned to a transition. It specifies the condition to enable or disable a transition and can use the entire state of the net rather than just the number of tokens in places [6]. Reward function defines the reward rate for each tangible marking of Petri Net based on which various quantitative measures can be done in the Net level. Marking dependent firing rate allows using the number of tokens in a chosen place multiplied by the basic rate of the transition. SRN model has the following elements: Finite set of the place (drawn as circles), Finite set of the transition defined as either a timed transition (drawn as thick transparent bar) or a immediate transition (drawn as thick black bar), set of the arc connecting the place and transition, multiplicity associated with the arc, and marking that denotes the number of token in each place.

Before introducing the model translation rules, different types of collaboration roles as reusable basic building blocks are demonstrated with the corresponding SRN model in Table I that can be utilized to form the collaborative building blocks.

The rules are the following:

Rule 1

The SRN model of a collaboration (Fig. 7), where collaboration connects only two collaboration roles, is formed by combining the basic building blocks type 2 and type 3 from Table I. Transition ***t*** in the SRN model is only realized by the overhead cost if service components A and B deploy on the same physical node as in this case, communication cost = 0, otherwise ***t*** is realized by both the communication & overhead cost.

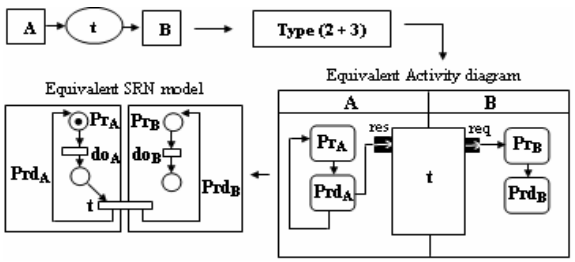


Figure 7. Graphical representation of rule 1

In the same way, SRN model of the collaboration can be demonstrated where the starting of the execution of the SRN model of collaboration role A depends on the token received from the external source.

Rule 2

For a composite structure, when a collaboration role A connects with *n* collaboration roles by *n* collaborations like a star graph (where n > 1) where each collaboration connects only two collaboration roles, the SRN model is formed by combining the basic building block of Table I which is shown in Fig. 8. In the first diagram of Fig. 8, if component A contains its own token, equivalent SRN model of the collaboration role A will be formed using basic building block type 1 from Table I. The same applies to the component B and C in the second diagram in Fig. 8.

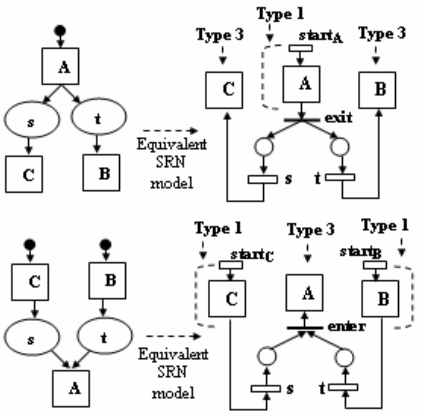


Figure 8. Graphical representation of rule 2

STM can be translated into a SRN model by converting each state into place and each transition into a timed transition with input/output arcs which is reflected in the transformation Rule 3.

Rule 3

Rule 3 demonstrates the equivalent SRN model of the STM of hardware and software components which are shown in the Fig. 9.

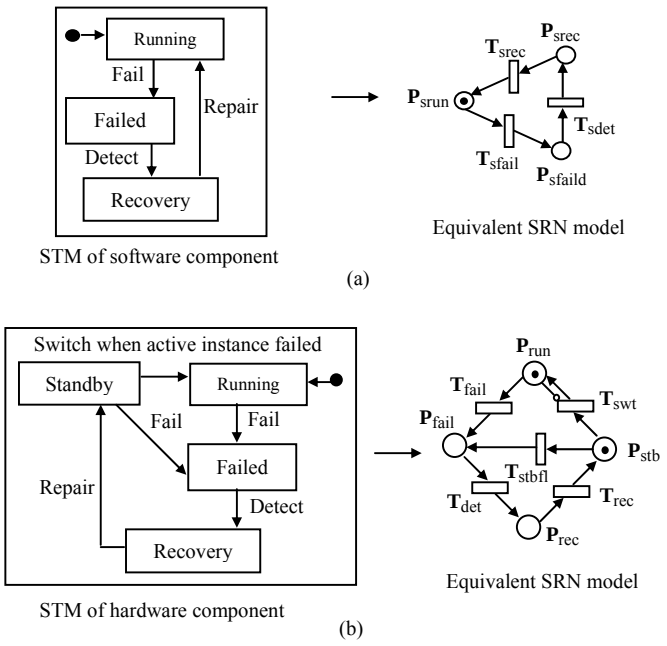


Figure 9 (a) SRN of Software process (b) SRN of hardware component

The SRN model for hardware component is shown in Fig. 9(b). A token in the place P_{run} represents the active hardware component and a token in P_{stb} represents a hot standby hardware component. When the transition T_{fail} fires, the token in P_{run} is removed and the transition T_{swt} is enabled. By the T_{swt}, which represents the failover, hot standby hardware component becomes an active component.

VIII. MODEL SYNCHRONIZATION

The model synchronization is achieved hierarchically which is illustrated in Fig. 10. Performance SRN is dependent on the dependability SRN. Transitions in dependability SRN may change the behavior of the performance SRN. Moreover, transitions in the SRN model for the software process also depend on the transitions in the SRN model of the hardware component. These dependencies in the SRN models are handled through model synchronization by incorporating guard functions [6].

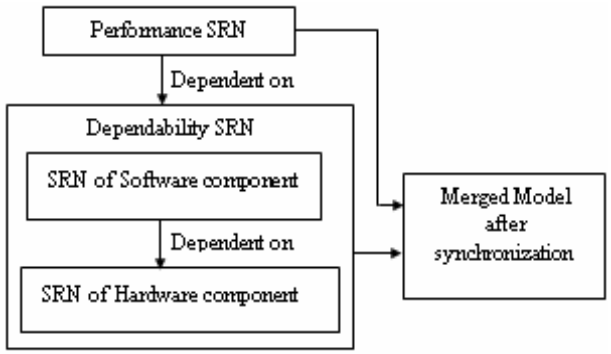


Figure 10. Model synchronization hierarchy

The model synchronization is focused in detail below:

A. Synchronization between the dependability SRN models in the dependability modeling layer

SRN model for the software process (Fig. 9(a)) is expanded by incorporating one additional place P_{hf}, three immediate transitions t_{hf}, t_{hsfl}, t_{hfr}, and one timed transition T_{recv} to synchronize the transitions in the SRN model for the software process with the SRN model for the hardware component. The expanded SRN model (Fig. 11(a)) is

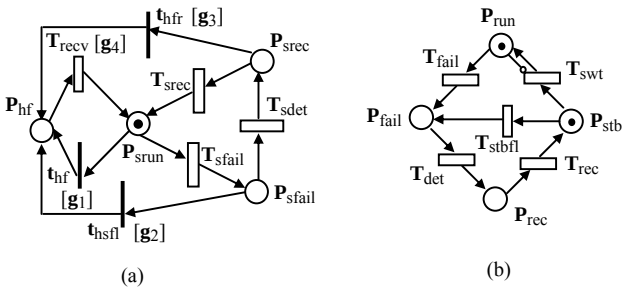


Figure 11. (a) Synchronized transition in the SRN model of the software process with the (b) SRN model of the hardware component

associated with four additional arcs such as $(P_{sfail} \times t_{hsfl}) \cup (t_{hsfl} \times P_{hf})$, $(P_{srec} \times t_{hfr}) \cup (t_{hfr} \times P_{hf})$, $(P_{srun} \times t_{hf}) \cup (t_{hf} \times P_{hf})$ and $(P_{hf} \times T_{recv}) \cup (T_{recv} \times P_{srun})$. The immediate transitions t_{hf}, t_{hsfl}, t_{hfr} will be enabled only when the hardware node (in Fig. 11 (b)) fails as failure of hardware node will stop operation of the software process. The timed transition T_{recv} will be enabled only when the hardware node will again start working after being recovered from failure. Four guard functions g_1, g_2, g_3, g_4 allow the four additional transitions t_{hf}, t_{hsfl}, t_{hfr} and T_{recv} of software process to work consistently with the change of states of the hardware node. The guard functions definitions are given in the Table II.

TABLE II. GUARD FUNCTIONS DEFINITION

Function	Definition
g_1, g_2, g_3	if (# P_{run} = = 0) 1 else 0
g_4	if (# P_{run} = = 1) 1 else 0

B. Synchronization between the dependability SRN & performance SRN

In order to synchronize the collaboration role activity, performance SRN model is expanded by incorporating one additional place P_{fl} and one immediate transition f_A shown in Fig. 12. After being deployed when collaboration role "A" starts execution, a checking will be performed to examine whether both software and hardware components are running or not. If both the components work the timed transition do_A will fire which represents the continuation of the execution of the collaboration role A. But if software resp. hardware components fail the immediate transition f_A will be fired which represents the quitting of the operation of collaboration role A. Guard function gr_A allows the immediate transition f_A to work consistently with the change of states of the software and hardware components.

Performance SRN model of parallel execution of collaboration roles are expanded by incorporating one additional place P_{fl} and immediate transitions f_{BC}, w_{BC} shown in Fig. 12. In our discussion, during the synchronization of the parallel processes it needs to ensure that failure of one process eventually stops providing service to the users. This could be achieved by immediate transition f_{BC}. If software resp. hardware components (Fig. 11) fail immediate transition f_{BC} will be fired which symbolizes the quitting of the operation of both parallel processes B and C rather than stopping either process B or C, thus postponing the execution of the service. Stopping only either the process B or C will result in inconsistent execution of the whole SRN and produce erroneous result. If both software and hardware components work fine the timed transition w_{BC} will fire to continue the execution of parallel processes B and C. Guard functions gr_{BC}, grw_{BC} allow the immediate

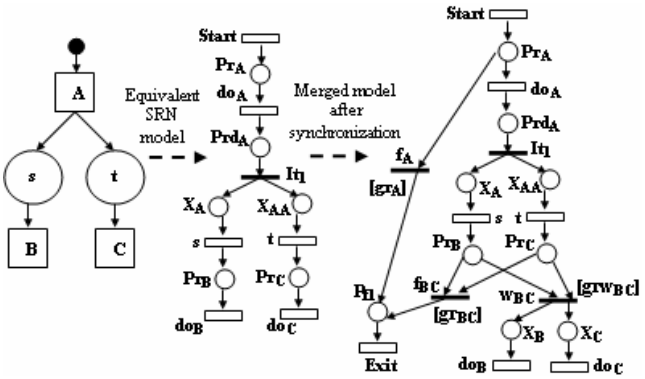


Figure 12. Synchronize the performance SRN model with dependability SRN

transition $f_{BC,}$ w_{BC} to work consistently with the change of the states of the software and hardware components. The guard function definitions are shown in the Table III.

Algorithms for model transformation rules and model synchronization process have been mentioned in Appendix A.

TABLE III. GUARD FUNCTIONS DEFINITION

Function	Definition
gr_A, gr_{BC}	if (# P_{srun} = = 0) 1 else 0
grw_{BC}	if (# P_{srun} = = 1) 1 else 0

IX. HIERARCHICAL MODEL FOR MTTF CALCULATION

System is composed of different types of hardware devices such as CPU, memory, storage device, cooler. Hence, to model the failure behavior of a hardware node absolutely, we need to consider failure behavior of all the hardware devices. But it is very demanding and not efficient with respect to execution time to consider behavior of all the hardware components during the SRN model generation. SRN model becomes very cumbersome and inefficient to execute. In order to solve the problem, we evaluate the mean time to failure (MTTF) of system using the hierarchical model in which a fault tree is used to represent the MTTF of the system by considering MTTF of every hardware component in the system. Later on, we consider this MTTF of the system in our dependability SRN model for hardware components (Fig. 9(b)) rather than considering failure behavior of all the hardware components individually. The below Fig. 13 introduces one example scenario of capturing failure behavior of the hardware components using fault tree where system is composed of different hardware devices such as one CPU, two memory interfaces, one storage device and one cooler. The system will work when CPU, one of the memory interfaces, storage device and cooler will run. Failure of both memory interfaces or failure of either CPU or storage device or cooler will result in the system unavailability.

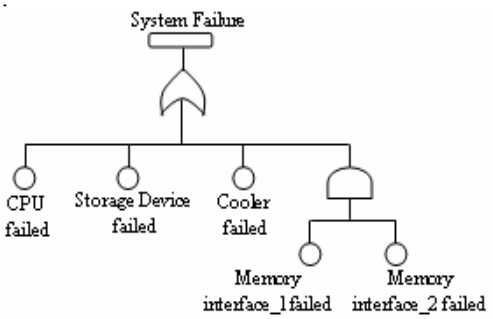


Figure 13. Fault tree model of System Failure

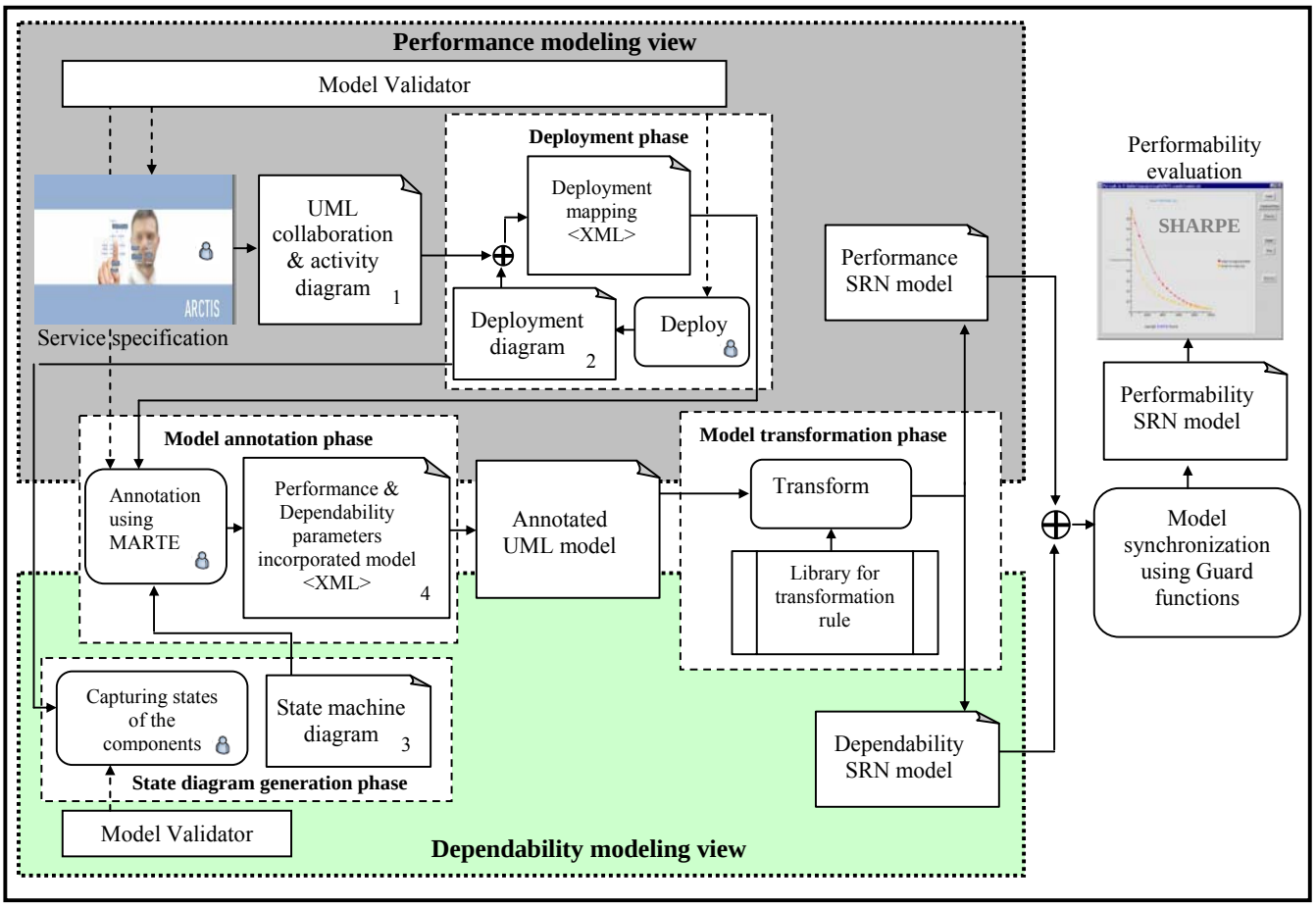


Figure 14. Tool support of our performability modeling framework

X. TOOL BASED SUPPORT OF THE PERFORMABILITY MODELING FRAMEWORK

The theoretical foundation of the approach is described in details in the above sections. We highlight the tool support of our performability modeling framework in Fig. 14. The partial input model of our framework is generated using Arctis tool which is integrated as plug-in into the eclipse IDE. In the evaluation side, SHARPE tool is used. We generate the annotated UML model from the UML collaboration diagram, deployment diagram, STM diagram, and the performance and dependability related parameters. From Fig. 14, it is evident that we need to define 4 inputs accordingly: in the performance modeling view, the first input UML collaboration diagram and the detail behavior of collaborative building block will be generated using the GUI (Graphical User Interface) editor of Arctis tool which will be saved as XML file and the other two inputs of performance modeling view will be generated as XML file such as deployment diagram and performance attributes incorporated UML model after deployment mapping. The inputs of the dependability modeling view such as STM diagram and dependability attributes incorporated UML model will be generated as XML file as well. We also define one output file in text format which is generated as a result of the model annotation phase denoting the annotated UML model. The annotated UML model file is then further used as an input for the model transformation phase to achieve automation in model transformation. In the model transformation phase, we automate the transformation process from annotated UML model to the SRN performability model following the model transformation rules and afterwards, merging of SRN performance and dependability model using guard functions. The input files are specified in XML formats. This is because of the fact that XML gives benefits to guarantee the robustness, flexibility to extend the existing file, and data validation. The output files are all in text format as the SHARPE tool, that evaluates the performance of the system, accepts the input as text format.

XI. CASE STUDY

As a representative example, we consider a scenario dealing with heuristically clustering of modules and assignment of clusters to nodes [17]. This scenario is sufficiently complex to show the applicability of our performability framework. The problem is defined in our approach as collaboration of $E = 10$ service components or collaboration roles (labeled $C_1 \ldots C_{10}$) to be deployed and $K = 14$ collaborations between them illustrated in Fig. 15. We consider three types of requirements in this specification. Besides the execution cost, communication cost, and cost for running background process, we have a restriction on components C_2, C_7, C_9 regarding their location. They must be bound to nodes $\boldsymbol{n_2}$, $\boldsymbol{n_1}$, $\boldsymbol{n_3}$ respectively. In this scenario, new service is generated by integrating and combining the existing service components that will be delivered conveniently by the system. For example, one new service is

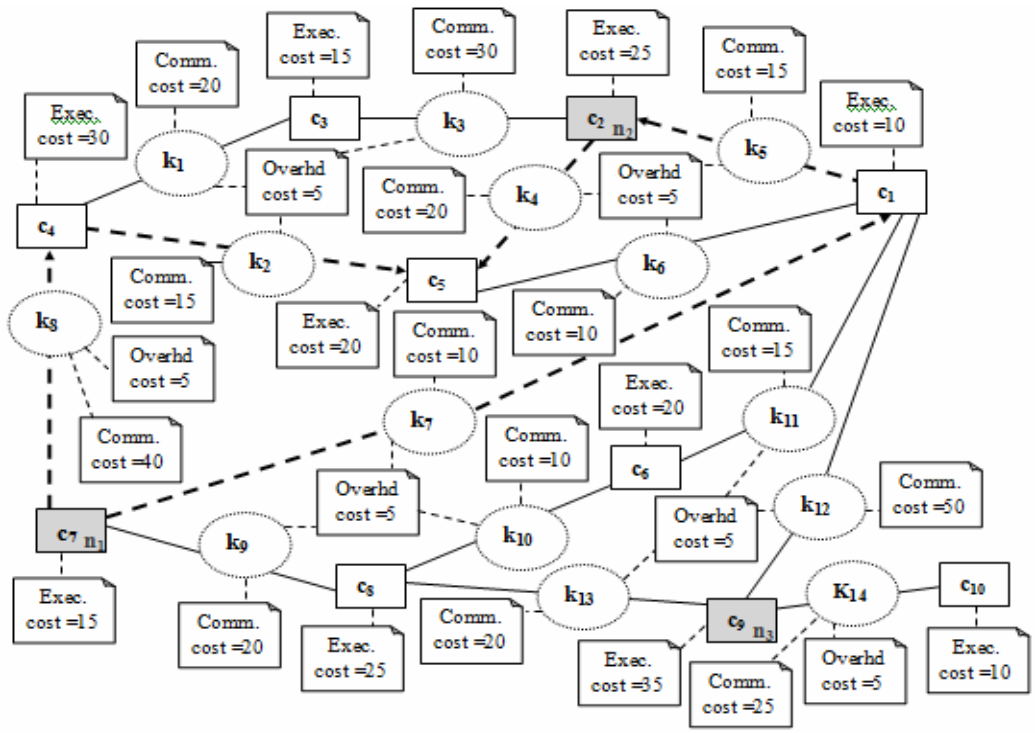


Figure 15. Collaboration & Components in the example Scenario

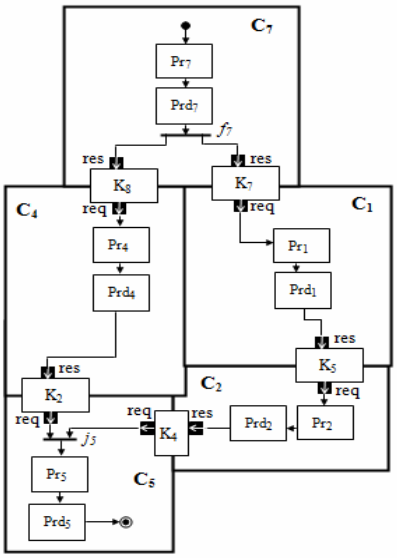


Figure 16. Composition of collaboration

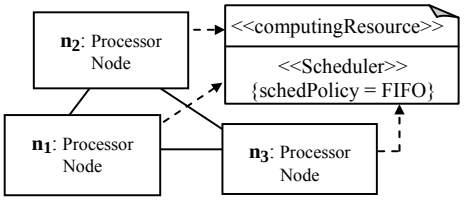


Figure 17. The target network of hosts

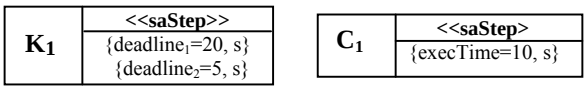


Figure 18. Annotated UML model

composed by combining the service components C_1, C_2, C_4, C_5, C_7 shown in Fig. 15 as thick dashed line. The internal behavior of the collaboration K_i is realized by the call behavior actions through the same UML activity diagram already demonstrated in Fig. 3(b). The composition of the collaboration role C_i of the delivered service by the system is demonstrated in Fig. 16. The initial node (●) indicates the starting of the activity. After being activated, each participant starts its processing of request which is mentioned by call behavior action Pr_i (Processing of the ith service component). Completions of the processing by the participants are mentioned by the call behavior action Prd_i (Processing done of the ith service component). The activity is started from the component C_7 where the semantics of the activity is realized by the token flow. After completion of the processing of the component C_7, the response is divided into two flows which are shown by the fork node f_7. The flows are activated towards component C_1 and C_4. After getting the response from the component C_1, processing of the components C_2 will be started. The response and request are mentioned by the streaming pin *res* and *req*. The processing of the component C_5 will be started after getting the responses from both component C_4 and C_2 which is realized by the join node j_5. After completion of the processing of component C_5, the activity is terminated which is mentioned by the end node (◉).

In this example, the target environment consists of $N = 3$ identical, interconnected nodes with no failure of network link, with a single provided property, namely processing power, and with infinite communication capacities shown in Fig. 17. The optimal deployment mapping can be observed in Table IV. The lowest possible deployment cost, according to equation (4) is: 17 + 100 + 70 = 187.

In order to annotate the UML diagrams in Fig. 16 and 17, we use the stereotypes *<<saStep>>* *<<computingResource>>*, *<<scheduler>>* and the tagged values *execTime*, *deadline* and *schedPolicy* which are already explained in section 5. Collaboration K_i (Fig. 18) is associated with two instances of *deadline* as collaborations in example scenario are associated with two kinds of cost:

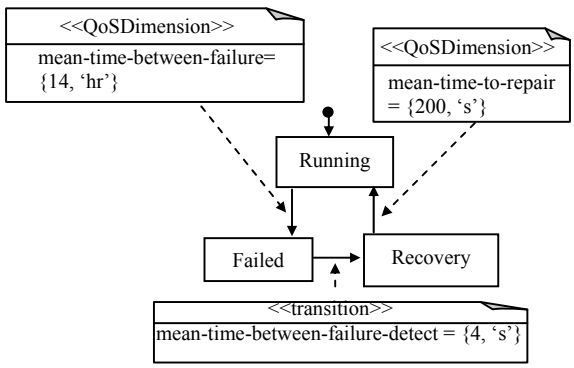


Figure 19. Annotated STM diagram of software component

TABLE IV. OPTIMAL DEPLOYMENT MAPPING

Node	Components	$\widehat{l}_n$	$\lvert \widehat{l}_n - T \rvert$	Internal collaborations
n_1	c_4, c_7, c_8	70	2	k_8, k_9
n_2	c_2, c_3, c_5	60	8	k_3, k_4
n_3	c_1, c_6, c_9, c_{10}	75	7	k_{11}, k_{12}, k_{14}
$\sum$ cost			17	100

communication cost and cost for running background process (BP). In order to annotate the STM UML diagram of software process (shown in Fig. 19), we use the stereotype *<<QoSDimension>>*, *<<transition>>* and attributes *mean-time-between-failures, mean-time-between-failure-detect* and *mean-time-to-repair* which are already mentioned in section VI. Annotation of the STM of hardware component can be demonstrated in the same way as STM of software process.

By considering the specification of reusable collaborative building blocks, deployment mapping, and the model transformation rule, the corresponding SRN model of our example scenario is illustrated in Fig. 20. In our discussion we consider M/M/1/n queuing system so that at most *n* jobs can be in the system at a time [3]. For generating the SRN model, firstly, we will consider the starting node (●). According to rule 1, it is represented by timed transition (denoted as *start*) and the arc connects to place Pr_7 (states of component C_7). When a token is deposited in place Pr_7, immediately a checking is done about the availability of both software and hardware components by inspecting the corresponding SRN models shown in Fig. 11. The availability of software and hardware components allows the firing of timed transition t_7 mentioning the continuation of the further execution. Otherwise, immediate transition f_7 will be fired mentioning the ending of the further execution because of software resp. hardware component failure. The enabling of immediate transition f_7 is realized by the guard function $\boldsymbol{gr_7}$. After the completion of the state transition from Pr_7 to Prd_7 (states of component C_7), immediately, the flow is divided into two branches (denoted by the immediate transition It_1) according to model transformation rule 2 (Fig. 8). The token is passed to place Pr_1 (states of component C_1) and Pr_4 (states of component C_4) after the firing of transitions K_7 and K_8. According to rule 1, collaboration K_8 is realized only by overhead cost as C_4 and C_7 deploy on the same processor node $\boldsymbol{n_1}$ (Table IV). The collaboration K_7 is realized both by the communication cost and overhead cost as C_1 and C_7 deploy on the two different nodes $\boldsymbol{n_3}$ and $\boldsymbol{n_1}$ (Table IV). When a token is deposited into place Pr_1 and Pr_4, immediately, a checking is done about the availability of both software and hardware components by inspecting the corresponding dependability SRN models illustrated in Fig. 11. The availability of software and hardware components allows the firing of immediate transition w_{14} which eventually enables the firing of timed transition t_1 mentioning the continuation of the further execution. The enabling of immediate transition w_{14} is realized by the guard function $\boldsymbol{grw_{14}}$. Otherwise, immediate transition f_{14} will be fired mentioning the ending of the further execution because of software resp. hardware component failure. The enabling of immediate transition f_{14} is realized by the guard function $\boldsymbol{gr_{14}}$. After the completion of the state transition from Pr_1 to Prd_1 (states of component C_1) the token is passed to Pr_2 (states of component C_2) according to rule 1, where timed transition K_5 is realized both by the communication and overhead cost. When a token is deposited into place Pr_2, immediately a checking is done about the availability of both software and hardware components by inspecting the corresponding dependability SRN models shown in Fig. 11. The availability of software and hardware components allows the firing of the immediate transition w_{24} which eventually enables the firing of timed transition t_2 and t_4 mentioning the continuation of the further execution. The enabling of immediate transition w_{24} is realized by the guard function $\boldsymbol{grw_{24}}$. Otherwise, immediate transition f_{24} guided by guard function $\boldsymbol{gr_{24}}$ will be fired mentioning the ending of the further execution because of software resp. hardware component failure. Afterwards, the merging of the result is realized by the immediate transition It_2 following the firing of transitions K_2 and K_4. Collaboration K_2 is realized both by the overhead cost and communication cost as C_4 and C_5 deploy on the different processor nodes $\boldsymbol{n_1}$ and $\boldsymbol{n_2}$ (Table IV). K_4 is replaced by the timed transition which is realized by the overhead cost as C_2 and C_5 deploy on the same node $\boldsymbol{n_2}$ (Table IV). When a token is deposited in place Pr_5 (state of component C_5), immediately, a checking is done about the availability of both software and hardware components by inspecting the corresponding SRN models illustrated in Fig. 11. The availability of software and hardware components allows the firing of timed transition t_5 mentioning the

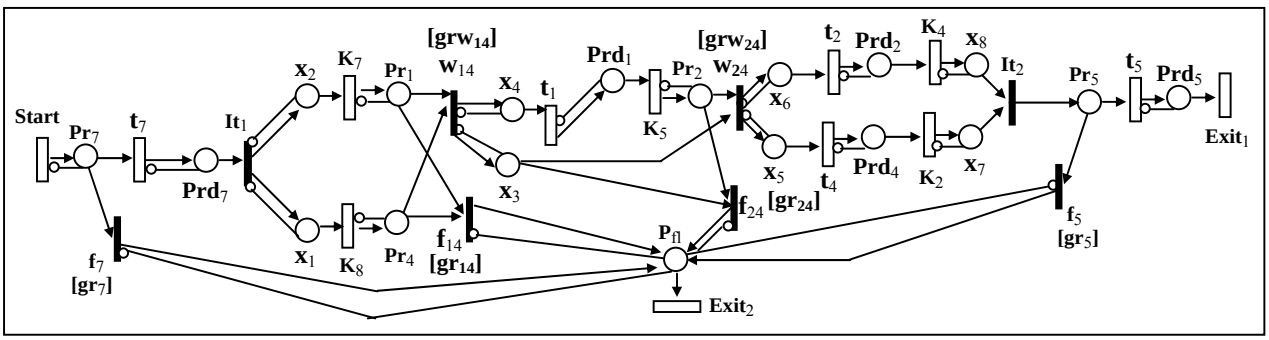


Figure 20. Equivalent SRN model of the example service

continuation of the further execution. Otherwise, immediate transition f_5 will be fired mentioning the ending of the further execution because of software resp. hardware component failure and the ending of the execution of the SRN model is realized by the timed transition $Exit_2$. The enabling of immediate transition f_5 is realized by the guard function $\boldsymbol{gr_5}$. After the completion of the state transition from Pr_5 to Prd_5 (states of component C_5) the ending of the execution of the SRN model is realized by the timed transition $Exit_1$. The definitions of guard functions $\boldsymbol{gr_7}$, $\boldsymbol{grw_{14}}$, $\boldsymbol{gr_{14}}$, $\boldsymbol{grw_{24}}$, $\boldsymbol{gr_{24}}$, and $\boldsymbol{gr_5}$ are mentioned in Table V, which is dependent on the execution of the SRN model of the corresponding STM of software and hardware instances illustrated in Fig. 11.

TABLE V. GUARD FUNCTIONS DEFINITION

Function	Definition
gr_7, gr_{14}, gr_{24}, gr_5	if (# P_{srun} = = 0) 1 else 0
grw_{14}, grw_{24}	if (# P_{srun} = = 1) 1 else 0

We use SHARPE [16] to execute the obtained synchronized SRN model and calculate the system's throughput and job success probability against failure rate of system components. Graphs in Fig. 21 show the throughput and job success probability of the system against the changing of the failure rate (sec^{-1}) of hardware and software components in the system.

XII. CONCLUSION AND FUTURE WORK

We presented a novel approach for model based performability evaluation of a distributed software system. The approach spans from system's dynamics demonstration through UML diagram as reusable building blocks to efficient deployment of service components in a distributed manner focusing on the QoS requirements. The main advantage of using the reusable software components allows the cooperation among several software components to be reused within one self-contained, encapsulated building block. Moreover, reusability thus assists in creating the distributed software systems from existing software components rather than developing the system from scratch which in turn facilitates the improvement of productivity and quality in accordance with the reduction in time and cost. We put emphasis to establish some important concerns relating to the specification and solution of performability models emphasizing the analysis of the system's dynamics. We design the framework in a hierarchical and modular way which has the advantage of introducing any modification or adjustment at a specific layer in a particular submodel rather than in the combined model according to any change in the specification. Among the important issues that come up in our development are flexibility of capturing the system's dynamics using our new reusable specification of building blocks, ease of understanding the intricacy of combined model generation, and evaluation from that specification by proposing model transformation. However, our eventual goal is to develop support for runtime redeployment of components, this way keeping the service within an allowed region of parameters defined by the requirements. As a result, with our proposed framework we can show that our logic will be a prominent candidate for a robust and adaptive service execution platform. The special property of SRN model like guard function keeps the performability model simpler by applying logical conditions that can be expressed graphically using input and inhibitor arcs which are limited by the following semantics: a logical "AND" for input arcs (all the input conditions must be satisfied), a logical "OR" for inhibitor arcs (any inhibitor condition is sufficient to disable the transition) [18]. However, the size of the underlying reachability set to generate a SRN model is major limitation for large and complex systems. Further work includes tackling the state explosion problems of reachability marking for large distributed systems. In addition, developing GUI editor is another future direction to generate UML deployment and state diagram and to incorporate performability related parameters. The plug-ins can be integrated into the Arctis tool which will provide the automated and incremental model checking while conducting model transformation.

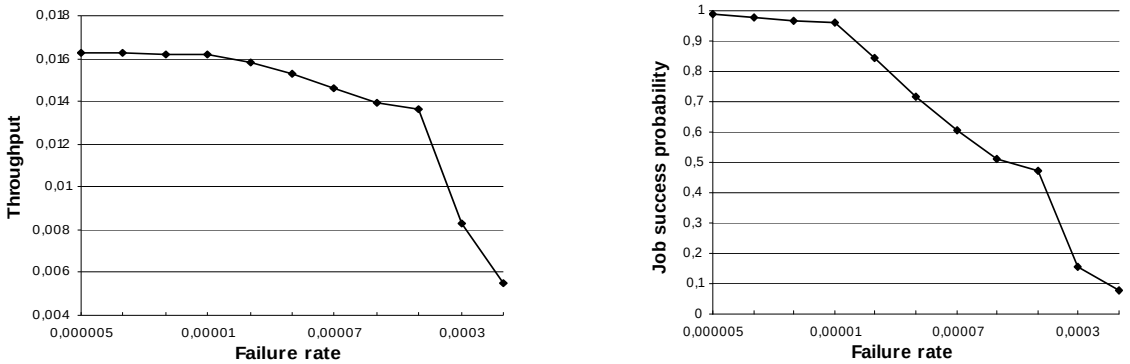


Figure 21. Numerical result of our example scenario

REFERENCES

[1] R H Khan, F Machida, P. Heegaard, and K S Trivedi, "From UML to SRN: A performability modeling framework considering service components deployment", Proceeding of the ICNS, pp. 118-127, IARIA, 2012

[2] F. A. Jawad and E. Johnsen, "Performability: the vital evaluation method for degradable systems and its most commonly used modeling method, Markov reward modeling", http://www.doc.ic.ac.uk/~nd/surprise_95/journal/vol4/eaj2/report.html, <retrieved May 2011>

[3] E. de Souza e. Silva, and H. R. Gali, "Performability analysis of computer systems: from model specification to solution", Performance evaluation 14, pp. 157-196, 1992

[4] K. S. Trivedi, "Probability and Statistics with Reliability, Queuing and Computer Science application", Wiley-Interscience publication, ISBN 0-471-33341-7, 2001

[5] OMG 2009, "OMG UML Superstructure", Version-2.2

[6] G. Ciardo, J. Muppala, and K. S. Trivedi, "Analyzing concurrent and fault-tolerant software using stochastic reward nets", Journal of Parallel and Distributed Computing, Vol. 15, 1992

[7] M. Csorba, P. Heegaard, and P. Herrmann, "Cost-Efficient Deployment of Collaborating Components", Proceedings of the DAIS, pp. 253–268, Springer, 2008

[8] OMG 2009, "UML Profile for MARTE: Modeling & Analysis of Real-Time Embedded Systems", V – 1.0

[9] N. Sato and Trivedi, "Stochastic Modeling of Composite Web Services for Closed-Form Analysis of Their Performance and Reliability Bottlenecks", Proceedings of the ICSOC, pp. 107-118, Springer, 2007

[10] P. Bracchi, B. Cukic, and Cortellesa, "Performability modeling of mobile software systems", Proceedings of the ISSRE, pp. 77-84, 2004

[11] N. D. Wet and P. Kritzinger, "Towards Model-Based Communication Protocol Performability Analysis with UML 2.0", http://pubs.cs.uct.ac.za/archive/00000150/01/No_10, <retrieved May 2011>

[12] Gonczy, Deri and Varro, "Model Driven Performability Analysis of Service Configurations with Reliable Messaging", Proceedings of the MDWE, 2008

[13] OMG 2009, "UML Profile for Modeling Quality of Service & Fault Tolerance Characteristics Specification", V-1.1

[14] R. H. Khan and P. Heegaard, "A Performance modeling framework incorporating cost efficient deployment of multiple collaborating components", Proceedings of the ICSECS, pp. 31-45, Springer, 2011

[15] F. A. Kramer, "ARCTIS", Department of Telematics, NTNU, http://arctis.item.ntnu.no, <retrieved May 2011>

[16] K. S. Trivedi and R. Sahner, "Symbolic Hierarchical Automated Reliability / Performance Evaluator (SHARPE)", Duke University, NC, 2002

[17] Mate J. Csorba, "Cost efficient deployment of distributed software services", PhD Thesis, NTNU, Norway, 2011

[18] Muppala, Ciardo and K. Trivedi, Stochastic reward nets for reliability prediction", Communications in Reliability, Maintainability and Serviceability, SAE International, 1994

[19] P. Herrmann and H. Krumm, "A Framework for Modeling Transfer Protocols", Computer Networks, Vol - 34, No - 2, pp.317–337, 2000

[20] Vidar Slåtten, "Model Checking Collaborative Service Specifications in TLA with TLC", Project Thesis, Norwegian University of Science and Technology, Trondheim, Norway, August 2007

[21] Lamport, "Specifying Systems", Addison-Wesley, 2002

[22] R. H. Khan and Poul E. Heegaard, "Software Performance evaluation utilizing UML Specification and SRN model and their formal representation", Submitted to a journal for reviewing.

[23] F. Kræmer, P. Herrmann, and R. Bræk, "Aligning UML 2 state machines & temporal logic for the effecient execution of services", Proceedings of the DOA, Springer, 2006

APPENDIX A

Algorithm 1: rule_1 (ExecCost, CommCost, Ovrhdcost, Mappings, CollaborationRoles)

```
1   If CollaborationRoles A self token generator then
2       Places += "Pr_A 1"
3   else (A has a external token generator)
4       Places += "Pr_A 0"
5   Places += "Prd_A 0"
6   Places += "Pr_B 0"
7   Places += "Prd_B 0"
8   Timed_Transitions += "do_A ind" + 1/ execution cost for collaborationRole A
9   Timed_Transitions += "do_B ind" + 1/ execution cost for collaboration role B
10  Timed_Transitions += "exit ind" + 1/ rate for the end transition
11  If CollaborationRoles A and B are deployed on the same node then
12      Timed_Transitions += "t ind" + 1/ overhead cost
13  else
14      Timed_Transitions += "t ind" + 1/ (overhead cost + communication cost)
15  If CollaborationRole A has a external token generator then
16      Timed_Transitions += "Start ind" + 1/ rate of the token generator
17  Inhibitor_Arcs += "Pr_A Start 1"
18  Inhibitor_Arcs += "Prd_A do_A 1"
19  Inhibitor_Arcs += "Pr_B t 1"
20  Inhibitor_Arcs += "Prd_B do_B 1"
21. Input_Arcs += "Pr_A do_A 1"
22  Input_Arcs += "Prd_A t 1"
23  Input_Arcs += "Pr_B do_B 1"
24  Input_Arcs += "Prd_B exit 1"
25  Output_Arcs += "do_A Prd_A 1"
26  Output_Arcs += "do_B Prd_B 1"
27  Output_Arcs += "t Pr_B 1"
28  if CollaborationRole A self token generator then
29      Output_Arcs += "t PrA 1"
30  else
31      Output_Arcs += "Start PrA 1"
32  Print Places, Timed_Transitions, Input_Arcs, Output_Arcs, Inhibitor_Arcs
33  return
```

Algorithm 2: rule_2_a (ExecCost, CommCost, Ovrhdcost, Mappings, CollaborationRoles)

```
1   Places += "Pr_A 0"
2   Places += "Prd_A 0"
3   Places += "Pr_B 0"
4   Places += "Prd_B 0"
5   Places += "Pr_C 0"
6   Places += "Prd_C 0"
7   Places += "X_c 0"
8   Places += "X_b 0"
9   Immediate_Transitions += "it ind 1"
10  Timed_Transitions += "Start ind" + 1 / rate of the
                         external token generator
11  Timed_Transitions += "do_A ind "+ 1 / execution cost
                         of collaboration role A
12  Timed_Transitions += "do_B ind "+ 1 / execution cost
                         of collaboration role B
13  Timed_Transitions += "doC ind "+ 1 / execution cost
                         of collaboration role C
14  if CollaborationRoles A and B are deployed on the
    same node then
15          Timed_Transitions += "t_B ind" + 1/ overhead
                                 cost
15  else
            Timed_Transitions += "t_B ind" + 1/ (overhead
                                 cost + communication cost)
16  if CollaborationRoles A and C are deployed on the same
    node then
17          Timed_Transitions += "t_C ind " + 1/ overhead
                                 cost
18  else
19          Timed_Transitions += "t_C ind"+1/ (overhead
                                 cost + communication cost)
20  Input_Arcs += "Pr_A do_A 1"
21  Input_Arcs += "Prd_A it 1"
22  Input_Arcs += "Pr_B do_B 1"
23  Input_Arcs += "Pr_C do_C 1"
24  Input_Arcs += "X_B t_B 1"
25  Input_Arcs += "X_C T_C 1"
26  Output_Arcs += "Start Pr_A 1"
27  Output_Arcs += "do_A Prd_A 1"
28  Output_Arcs += "it X_b 1"
29  Output_Arcs += "it X_c 1"
30  Output_Arcs += "t_B Pr_B 1"
31  Output_Arcs += "t_C Pr_C 1"
32  Output_Arcs += "do_B Prd_B 1"
33  Output_Arcs += "do_C Prd_C 1"
34  Inhibitor_Arcs += "Pr_A Start 1"
35  Inhibitor_Arcs += "Prd_A do_A 1"
36  Inhibitor_Arcs += "X_b it 1"
37  Inhibitor_Arcs += "X_c IT 1"
38  Inhibitor_Arcs += "Pr_B t_B 1"
39  Inhibitor_Arcs += "Pr_C t_C 1"
40  Inhibitor_Arcs += "Prd_B do_B 1"
41  Inhibitor_Arcs += "Prd_C do_C 1"
42  Print Places, Immediate_Transitions, Timed_Transitions,
    Input_Arcs, Output_Arcs, Inhibitor_Arcs
43  return
```

Algorithm 3: rule_2_b (ExecCost, CommCost, Ovrhdcost, Mappings, CollaborationRoles)

```
1   Places += "Pr_A 0"
2   Places += "Prd_A 0"
3   Places += "Pr_B 0"
4   Places += "Prd_B 0"
5   Places += "Pr_C 0"
6   Places += "Prd_C 0"
7   Places += "X_b 0"
8   Places += "X_c 0"
9   Immediate_Transitions += "it ind 1"
10  Timed_Transitions += "Start_B ind " + 1 / rate of the
                         external token generator for B
11  Timed_Transitions += "Start_C ind " + 1 / rate of the
                         external token generator for C
12  Timed_Transitions += "do_A ind " + 1 / execution cost
                         of CollaborationRoles A
13  Timed_Transitions += "do_B ind " + 1 / execution cost
                         of CollaborationRoles B
14  Timed_Transitions += "do_C ind " + 1 / execution cost
                         of CollaborationRoles C
15  if CollaborationRoles A and B are deployed on the same
    node then
16          Timed_Transitions += "t_B ind" + 1/ overhead
                                 cost
17  else
18          Timed_Transitions += "t_B ind" + 1/ (overhead
                                 cost + communication cost)
19  if CollaborationRoles A and C are deployed on the same
    node then
20          Timed_Transitions += "t_C ind" + 1/ overhead
                                 cost
21. else
22          Timed_Transitions += "t_C ind" + 1/ (overhead
                                 cost + communication cost)
23  Input_Arcs += "Pr_A do_A 1"
24  Input_Arcs += "Pr_B do_B 1"
25  Input_Arcs += "Prd_B t_B 1"
26  Input_Arcs += "X_b it 1"
27  Input_Arcs += "Pr_C do_C 1"
28  Input_Arcs += "Prd_C t_C 1"
29  Input_Arcs += "Xc it 1"
30  Output_Arcs += "it Pr_A 1"
31  Output_Arcs += "do_A Prd_A 1"
32  Output_Arcs += "Start_B Pr_B 1"
33  Output_Arcs += "do_B Prd_B 1"
34  Output_Arcs += "t_B X_b 1"
35  Output_Arcs += "Start_C Pr_C 1"
36  Output_Arcs += "do_C Prd_C 1"
37  Output_Arcs += "t_C X_c 1"
38  Inhibitor_Arcs += "Pr_B Start_B 1"
39  Inhibitor_Arcs += "Prd_B do_B 1"
40  Inhibitor_Arcs += "X_b t_B 1"
41  Inhibitor_Arcs += "Pr_C Start_C 1"
42  Inhibitor_Arcs += "Prd_C do_C 1"
43  Inhibitor_Arcs += "X_c t_C 1"
45  Inhibitor_Arcs += "Pr_A it 1"
46  Inhibitor_Arcs += "Prd_A do_A 1"
47  Print Places, Immediate_Transitions, Timed_Transitions,
        Input_Arcs, Output_Arcs, Inhibitor_Arcs
48  return
```

Algorithm 4: rule_3_ hardware_srn()

```
1       Places += "H_run 1"
2       Places += "H_fail 0"
3       Places += "H_recover 0"
4       Places += "H_backup 1"
5       Timed_Transitions += "T_fl ind" + 1/ cost for the
                transition between H_run and H_fail
6       Timed_Transitions += "T_dt ind" + 1/ cost for the
                transition between H_fail and H_recover
7       Timed_Transitions += "T_rcv ind" + + 1/ cost for
                the transition between H_recover and H_backup
8       Timed_Transitions += "T_bfl ind" + 1/ cost for the
                transition between H_backup and H_fail
9       Timed_Transitions += "T_sw ind" + 1/ cost for the
                transition between H_backup and H_run
10      Input_Arcs += "H_run T_fl 1"
11      Input_Arcs += "H_fail T_dt 1"
12      Input_Arcs += "H_recover T_rcv 1"
13      Input_Arcs += "H_backup T_sw 1"
14      Input_Arcs += "H_backup T_bfl 1"
15      Output_Arcs += "T_fl H_fail 1"
16      Output_Arcs += "T_dt H_recover 1"
17      Output_Arcs += "T_rcv H_backup 1"
18      Output_Arcs += "T_sw H_run 1"
19      Output_Arcs += "T_bfl H_fail 1"
20      Inhibitor_Arcs += "H_run T_sw 1"
21      Print Places, Timed_Transitions, Input_Arcs,
            Output_Arcs, Inhibitor_Arcs
22  return
```

Algorithm 5: rule_3_software_srn()

```
1       Places += "S_run 1"
2       Places += "S_fail 0"
3       Places += "S_recover 0"
4       Timed_Transitions += "T_sfl ind " + 1/ cost for the
                transition between S_run and S_fail
5       Timed_Transitions += "T_sdt ind " + 1/ cost for the
                transition between S_fail and S_recovery
6       Timed_Transitions += "T_srcv ind " + + 1/ cost for
                the transition between S_recover and S_run
7       Input_Arcs += "S_run T_sfl 1"
8       Input_Arcs += "S_fail T_sdt 1"
9       Input_Arcs += "S_recover T_srcv 1"
10      Output_Arcs += "T_sfl S_fail 1"
11      Output_Arcs += "T_sdt S_recover 1"
12      Output_Arcs += "T_srcv S_backup 1"
13      Print Places, Timed_Transitions, Input_Arcs,
            Output_Arcs
14      return
```

Algorithm 6: software_sync_srn()

```
1       Places += "S_run 1"
2       Places += "S_fail 0"
3       Places += "S_recover 0"
4       Places += "P_hf 0"
5       Timed_Transitions += "T_sfl ind" + 1/ cost for the
                transition between S_run and S_fail
6       Timed_Transitions += "T_sdt ind" + 1/ cost for the
                transition between S_fail and S_recover
7       Timed_Transitions += "T_srcv ind" + 1/ cost for the
                transition between S_recover and S_run
8       Timed_Transitions += "T_recv ind" + 1/ cost for the
            transition between P_hf and S_run + "guard hd_up()"
9       Immediate_Transitions += "t_hfl ind 1 guard
                hd_down()"
10      Immediate_Transitions += "t_hf ind 1 guard
                hd_down()"
11      Immediate_Transitions += "t_hfr ind 1 guard
                hd_down()"
12      Input_Arcs += "S_run T_sfl 1"
13      Input_Arcs += "S_fail T_sdt 1"
14      Input_Arcs += "S_recover T_srcv 1"
15      Input_Arcs += "S_run t_hf 1"
16      Input_Arcs += "S_fail t_hfl 1"
17      Input_Arcs += "S_recover t_hfr 1"
18      Output_Arcs += "T_sfl S_fail 1"
19      Output_Arcs += "T_sdt S_recover 1"
20      Output_Arcs += "T_srcv S_run 1"
21      Output_Arcs += "t_hfl P_hf 1"
22      Output_Arcs += "t_hf P_hf 1"
23      Output_Arcs += "t_hfr P_hf 1"
24      Output_Arcs += "T_recv S_run 1"
25      Print Places, Timed_Transitions, Immediate_Transitions,
            Input_Arcs, Output_Arcs
26      return
```

hd_up()

```
1       if place H_run has one token then
2               return TRUE
3       else
4               return FALSE
5       return
```

hd_down()

```
1       if place H_run has zero token then
2               return TRUE
3       else
4               return FALSE
5   return
```

Algorithm 7: collaboration_role_sync_srn()

```
1     Places += "Pr_A 0"
2     Places += "Prd_A 0"
3     Places += "P_fl 0"
4     Immediate_Transitions += "f_A ind 1 guard sw_down()"
5     Timed_Transitions += "Start ind" + 1 / rate of the
                           external token generator
6     Timed_Transitions += "do_A ind" + 1 / execution cost
                           of collaboration role A
7     Timed_Transitions += "End_1 ind" + 1 / rate of the End_1
                           transition
8     Timed_Transitions += "End_2 ind" + 1 / rate of the End_2
                           transition
9     Input_Arcs += "Pr_A do_A 1"
10    Input_Arcs += "Pr_A f_A 1"
11    Input_Arcs += "Prd_A End_1 1"
12    Input_Arcs += "f_A End_2 1"
13    Output_Arcs += "Start Pr_A 1"
14    Output_Arcs += "do_A Prd_A 1"
15    Output_Arcs += "f_A P_fl 1"
16    Inhibitor_Arcs += "Pr_A Start 1"
17    Inhibitor_Arcs += "Prd_A do_A 1"
18    Inhibitor_Arcs += "P_fl f_A 1"
19    Print Places, Timed_Transitions, mmediate_Transitions,
          Input_Arcs, Output_Arcs, Inhibitor_Arcs
20    return
```

sw_down()

```
1     if place H_run has zero token then
2           return TRUE
3     else
4           return FALSE
5     return
```

Algorithm 8: building_block_sync_srn()

```
1     Places += "Pr_A 0"
2     Places += "Prd_A 0"
3     Places += "Pr_B 0"
4     Places += "Prd_B 0"
5     Places += "P_fl 0"
6     Immediate_Transitions += "f_A ind 1 guard
                               sw_down()"
7     Immediate_Transitions += "f_B ind 1 guard
                               sw_down()"
8     Timed_Transitions += "do_A ind" + 1 / execution
                           cost of collaboration role A
9     Timed_Transitions += "do_A ind" + 1 / execution
                           cost of collaboration role B
10    Timed_Transitions += "Start ind" + 1 / rate of
                           Start
11    if CollaborationRoles A and B are deployed on
      the same node then
12          Timed_Transitions += "T ind" + 1/
                                 overhead cost
13    else
14          Timed_Transitions += "T ind" + 1/
                (overhead cost + communication cost)
15    Input_Arcs += "Pr_A do_A 1"
16    Input_Arcs += "Pr_A f_A 1"
17    Input_Arcs += "Prd_A T 1"
18    Input_Arcs += "Pr_B do_B 1"
19    Input_Arcs += "Prd_B f_B 1"
20    Output_Arcs += "Start Pr_A 1"
21    Output_Arcs += "f_A P_fl 1"
22    Output_Arcs += "f_B P_fl 1"
23    Output_Arcs += "do_A Prd_A 1"
24    Output_Arcs += "do_B Prd_B 1"
25    Output_Arcs += "T Pr_B 1"
26    Inhibitor_Arcs += "Pr_A Start 1"
27    Inhibitor_Arcs += "Prd_A do_A 1"
28    Inhibitor_Arcs += "Pr_B T 1"
29    Inhibitor_Arcs += "PrdB doB 1"
30    Inhibitor_Arcs += "P_fl f_A 1"
31    Inhibitor_Arcs += "P_fl f_B 1"
32    Print Places, Timed_Transitions,
      Immediate_Transitions, Input_Arcs, Output_Arcs,
      Inhibitor_Arcs
33    return
```

sw_down()

```
1     if place S_run has zero token then
2           return TRUE
3     else
4           return FALSE
5     return
```

Algorithm 9: parallal_process_sync_srn()

```
1   Places += "Pr_A 0"
2   Places += "Prd_A 0"
3   Places += "Xa_1 0"
4   Places += "Xa_2 0"
5   Places += "P_fl 0"
6   Places += "Pr_B 0"
7   Places += "Prd_B 0"
8   Places += "Pr_C 0"
9   Places += "Prd_C 0"
10  Places += "X_B 0"
11  Places += "X_C 0"
12  Immediate_Transitions += "it ind 1"
13  Immediate_Transitions += "f_BC ind 1 guard sw_up()"
14  Immediate_Transitions += "f´_BC ind 1 guard sw_down()"
15. Timed_Transitions += "Start ind" + 1 / Start transition rate
16  if CollaborationRoles A and B are deployed on the same node then
17          Timed_Transitions += "T_B ind" + 1/ overhead cost
18  else
19          Timed_Transitions += "T_B ind" + 1/ (overhead cost + communication cost)
20  if CollaborationRoles A and C are deployed on the same node then
21          Timed_Transitions += "T_C ind" + 1/ overhead cost
22  else
23          Timed_Transitions += "T_C ind" + 1/ (overhead cost + communication cost)
24  Timed_Transitions += "End ind" + 1 / End transition rate
25  Input_Arcs += "Pr_A do_A 1"
26  Input_Arcs += "Prd_A it 1"
27  Input_Arcs += "Xa_1 T_B 1"
28  Input_Arcs += "Xa_2 T_C 1"
29  Input_Arcs += "Pr_B f_BC 1"
30  Input_Arcs += "Pr_B f´_BC 1"
31  Input_Arcs += "Pr_C f_BC 1"
32  Input_Arcs += "Pr_C f´_BC 1"
33  Input_Arcs += "P_fl End 1"
34  Input_Arcs += "X_B do_B 1"
35  Input_Arcs += "X_C do_C 1"
36  Output_Arcs += "Start Pr_A 1"
37  Output_Arcs += "do_A Prd_A 1"
38  Output_Arcs += "it Xa_1 1"
39  Output_Arcs += "it Xa_2 1"
40  Output_Arcs += "T_B Pr_B 1"
41  Output_Arcs += "T_C Pr_C 1"
42  Output_Arcs += "f_BC X_B 1"
43  Output_Arcs += "f_BC XC 1"
44  Output_Arcs += "f´_BC P_fl 1"
45  Output_Arcs += "do_B Prd_B 1"
46  Output_Arcs += "do_C Prd_C 1"
47  Inhibitor_Arcs += "Pr_A Start 1"
48  Inhibitor_Arcs += "Prd_A do_A 1"
49  Inhibitor_Arcs += "Xa_1 it 1"
50  Inhibitor_Arcs += "Xa_2 it 1"
51  Inhibitor_Arcs += "Pr_B T_B 1"
52  Inhibitor_Arcs += "Pr_C T_C 1"
53  Inhibitor_Arcs += "P_fl f´_BC 1"
54  Inhibitor_Arcs += "X_B f_BC 1"
55  Inhibitor_Arcs += "X_C f_BC 1"
56  Inhibitor_Arcs += "Prd_B do_B 1"
57  Inhibitor_Arcs += "Prd_C do_C 1"
58  Print Places, Timed_Transitions, Immediate_Transitions,
        Input_Arcs, Output_Arcs, Inhibitor_Arcs
59  return
```

sw_up()

```
1   if place S_run has one token then
2           return TRUE
3   else
4           return FALSE
5   return
```

sw_down()

```
1   if place S_run has zero token then
2           return TRUE
3   else
4           return FALSE
5   return
```

Algorithm 9: basic_bulding_block_srn()

```
1       if CollaborationRoles A has a self token generator then
2               Places += "Pr_i 1"
3       else
4               Places += "Pr_i 0"
5       Places += "Prd_i 0"
6       Timed_Transitions += "do ind " + 1/execution cost for
                                        collaboration role i
7       if i is getting token from external token generator then
8               Timed_Transitions += "Start ind" + 1 / Start
                                                        rate
9               Output_Arcs += "Start Pr_i 1"
10              Inhibitor_Arcs += "Pr_i Start 1"
11              Inhibitor_Arcs += "Prd_i do 1"
12      else if i is getting token from another CollaborationRoles
13              Timed_Transitions += "Enter ind" + 1 / cost of
                                                the  transition
14              Output_Arcs += "Enter Pr_i 1"
15      else
16              Output_Arcs += "Exit Pr_i 1"
17              Input_Arcs += "Pr_i do 1"
18      if i is passing its token then
19              Timed_Transitions += "Exit ind " + 1 / rate for
                                                        Exit
20              Input_Arcs += "Prd_i Exit 1"
21              Output_Arcs += "do Prd_i 1"
22      Print Places, Timed_Transitions, Input_Arcs, Output_Arcs,
                                                Inhibitor_Arcs
23      return
```

Using BGP to Reduce Power Consumption in Core and Edge Networks: A Metric-Based Approach

Shankar Raman*, Balaji Venkat†, and Gaurav Raina†
India-UK Advanced Technology Centre of Excellence in Next Generation Networks
*Department of Computer Science and Engineering, †Department of Electrical Engineering
Indian Institute of Technology Madras, Chennai 600 036, India
Email: mjsraman@cse.iitm.ac.in, balajivenkat@tenet.res.in, gaurav@ee.iitm.ac.in

***Abstract*—Power reduction methods at the device and the network levels in the Internet continue to attract attention. At the device level, power reduction is usually achieved by using low-power consuming devices or by switching off unused components. At the network level; re-engineering, reconfiguring, and re-routing of data packets can help in reducing power consumption. In this paper, we present a metric-based approach to route data packets through a low-power path from a source to a destination at the level of Autonomous Systems (AS). We propose that the Border Gateway Protocol (BGP) exchange a new attribute; namely, *consumed-power to available-bandwidth* of an AS with neighbouring AS. Using this proposed metric, the AS Border Routers can readily identify the low-power path from a source to a destination. We propose appropriate modifications to the BGPs path selection algorithm to include the low-power path criteria. We consider the effects that the proposed approach has on two parameters: (a) power reduction achieved as compared to the shortest path routing, and (b) the increase in path length from source to destination. The increase in the number of hops would be a consequence of re-routing through low-power AS. Simulations show that there could be significant gains in power reduction, if an increase in the number of hops is acceptable. Our work suggests that this trade off merits consideration as the power consumption of an AS is significant.**

***Index Terms*—Autonomous Systems, Border Gateway Protocol, Low-power Paths, Traffic Engineering.**

I. Introduction

Power reduction methods that aim to reduce energy consumption of the Internet have evoked much interest. In [1], a metric-based method for reducing the power consumption of the core and edge networks was proposed. This power reduction problem assumes significance as estimates predict a 300% increase, when access speeds move from 10 Mbps to 100 Mbps [2]. Numerous approaches have been proposed to reduce the power consumption ranging from designing low-power routers and switches, to optimizing the network topology using traffic engineering approaches [3].

Low-power router and switch design aim at reducing the power consumed by hardware components such as transmission link, lookup tables and memory. In [4], it is shown that the link power consumption can vary by 20 Watts from the base power, between idle and traffic scenarios. Hence, the authors suggest fully utilizing the line-card. The idea is that operating at full throughput will lead to less power per bit. Therefore, larger packet lengths will consume lower power.

The two important components that have received attention for high power consumption are static and dynamic RAM-based buffers (SRAM, DRAM) and Ternary Content Addressable Memories (TCAM). A 40 Gb/s line card would require more than 300 SRAM chips and consume 2.5 kW [5]. Some variants of TCAMs have been proposed for high speed lines with reduced power consumption [6]. But these schemes cannot scale forever. For some modeling work associated with buffer sizes, which can also lead to a reduction in power at the router architecture level, see [7], [8], and references therein.

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [9], a subset of IP router interfaces are put to sleep, using an Energy Aware Routing (EAR) after calculating shortest path trees of the network from each router. Such a technique is useful in setting up paths within an Autonomous System (AS). In [10], the authors provide a way to introduce hardware standby primitives and apply traffic engineering methods to coordinate and reduce power consumption under given network operational constraints. Power savings while switching from 1 Gbps to 100 Mbps is approximately 4 Watts and from 100 Mbps to 10 Mbps around 0.1 Watts. Hence, instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption. A detailed review on energy efficiency of the Internet is given in [11].

Multilayer traffic engineering based methods make use of parameters such as resource usage, bandwidth, throughput and Quality of Service (QoS) measures, for power reduction. In [12], an approach for reducing intra-AS power consumption for optical networks using Dijkstra's shortest path algorithm is proposed. The input assumes the existence of a network topology for constructing an auxiliary graph. This topology is easy to obtain for an intra-AS scenario. Traffic is then re-routed through the low-power optimized links.

The following issues exist in power reduction schemes today.

a A common method for reducing power consumption in the Internet is to switch unused devices and links to sleep state. Data packets are then routed through the functional links. This method is very localized and does not consider the power increase in the adjacent device that carries the extra traffic. Further this solution

is dependent on the underlying technology used by the devices. Also service level agreements between Internet Service Providers (ISPs) may be such that switching off the links may not be permissible.

b Power reduction schemes should not operate in isolation. They must be hierarchical so that they are applicable at various Internet hierarchies such as within the enterprise, or AS, and between AS. Further, there must not be any large variation in the algorithms implemented at the various levels of hierarchy. If multiple schemes are implemented at various levels of the hierarchy, then a way to coordinate these schemes become essential.

c Distributed solutions for power reduction have been used in adhoc wireless networks [13]. Such schemes may not be extensible to large networks such as the Internet. Further any proposed scheme must be extensible to multicast networks as well and not be limited to unicast.

Some of these drawbacks were addressed in our earlier scheme which was applicable for Multi-Protocol Label Switching (MPLS)-based networks [1]. MPLS label switched paths that traverse multiple AS carry traffic from a head-end to a tail-end AS, use Border Gateway Protocol (BGP) for exchanging routing and topology related information. In [1], the low-power path was detected by identifying the topology of the Internet. This topology at the AS level was obtained using the method presented in [14], where one of the attributes of BGP, AS-PATH-INFO, was used. The Constrained Shortest Path First algorithm (CSPF) uses this AS level topology with *consumed-power to available-bandwidth* (PWR) metric as a constraint, to determine the low-power path from the head-end to the tail-end. The PWR metric can be exchanged among the collaborating AS using BGP. It was shown that explicit routing can be achieved between the head and tail-ends through the low-power paths connecting the AS using inter-AS Traffic Engineered Label Switched Path (TE-LSP) that span multiple AS. However, this method has communication overhead in order to setup the path. These overheads occur in the form of information exchange between the entities in the network.

In order to avoid this, we propose modifications to the BGP path selection algorithm. This reduces the communication overhead associated with respect to setting up of the path. We introduce a new path selection rule to ensure that routing paths are established based on PWR metric by BGP rather than using inter-AS TE LSP. Simulations show that the PWR metric-based algorithms can lead to a power reduction which is as high as 70% over the conventional CSPF hop based variant. The power reduction obtained depends on the connectivity of the topology as well as the PWR metric distribution. There could be up to a 50% increase in the number of hops when compared with the shortest path algorithm. It has been suggested that for Internet Protocol (IP)-based networks such increase in hops may not have much impact in performance at the application layers [4].

The rest of the paper is organized in the following manner. In Section II, an overview of the BGP routing protocol as well as the inter-AS TE-LSP based algorithm is presented. Section III addresses ways to reduce the communication time complexity by proposing a method for establishing low-power paths using BGP path selection. Simulations are discussed in Section IV. In Section V, a brief discussion on the implementation and emulation using OpenFlow and Quagga is presented. A discussion on the comparison with our previously proposed implementation is presented in Section VI. We outline our contributions and highlight avenues for future research in Section VII.

II. Preliminaries

In this section, we present an introduction to the BGP protocol and the inter-AS TE-LSP based method for inter-AS power reduction.

A. Border Gateway Protocol

BGP [15] performs routing between multiple AS by exchanging routing and reachability information with other systems implementing BGP. BGP installs routing tables using a path selection algorithm. Routing information exchanges happen between multiple AS. This is classified under Exterior Border Gateway Protocol (eBGP). Internal BGP (iBGP) peering is used between the border routers in an AS and if necessary, between the core routers as well. Internal BGP expects a mesh topology. As such a network topology can become unmanageable due to scalability, route reflectors that re-advertise only the best path information is used to convey information to other iBGP routers.

BGP's routing decision is based on various static and dynamic parameters. Some examples for static parameters that affect routing decisions include multi-exit discriminator (MED) and local preference (LOCAL_PREF) values. Routing through oldest paths and AS-Path lengths are some examples for dynamic parameters. For a detailed discussion refer to [16], [17].

B. Inter-AS TE LSP power reduction using BGP

In our previous work, we presented a methodology for addressing the power reduction problem in the core and edge networks using BGP. The methodology was divided into four parts:

1) constructing the topology of the AS by a device,
2) assigning the PWR metric to the links connecting the AS,
3) calculating the low-power paths in the AS topology, and
4) establishing the path from source to the destination using traffic-engineering techniques.

We now briefly review the algorithms and discuss the computation and communication time complexity issues.

1) Constructing the network topology using BGP strands: The inter-AS topology can be modeled as a directed graph $G = (V, E, f)$ where the vertices (V) are mapped to AS and the edges (E) map the link that connect the neighboring AS. The direction (f) on the edge, represents the data flow from the head-end to the tail-end AS. To obtain the inter-AS topology, we use the approach from [14]. In this approach

a sub-graph of the Internet topology, can be obtained by collecting several prefix updates in BGP. This is illustrated in Figure 1 which shows the different graph strands of an AS recorded from the BGP packets.

Each vertex in this graph is assigned a weight according to the PWR metric of the AS, as seen from an AS Border Router (ASBR). Since there can be more than one ASBR associated with an AS, a vertex can have more than one PWR metric. Note that the ASBRs act as an entry point to the AS. Each of the PWR metrics for a vertex are assigned to the ingress links of the ASBRs. Figure 2 shows the merged strands forming the topology sub-graph where the weight of the vertices are mapped to the ingress edges. A reference AS level topology derived from 100 strands of AS-PATH-INFO received by an AS had 46 nodes with 15% connectivity in the topology. We define connectivity as a percentage of links present in the topology when compared with a complete graph of N nodes, which is $\frac{N(N-1)}{2}$.

2) PWR metric calculation: The numerator of the PWR metric is calculated for the AS at each ingress ASBR. We obtain the summation of power consumed at the major Provider (P) and Provider Edge (PE) routers within an AS. These can be obtained by using any of the intra-AS power calculation technique. The idea is to obtain the consumed-power of the AS which is the averaged consumed-power for all the routers within an AS. This value is divided by the maximum available-bandwidth at each of the ASBRs egress link. This step is necessary as the requested bandwidth for any path from the head-end to the tail-end using the ASBR is limited by the available-bandwidth in the ASBRs egress links. Note that Simple Network Management Protocol (SNMP) can also be used to extract this power information [18] offline.

The highest available bandwidth amongst the ASBRs egress links is used as the denominator in the PWR metric computation. Once the requested bandwidth is available, then consumed power plays a major role in determining the path from the source to the destination. PWR metric is used as a mapping function for each of the ingress link of the ASBR of an AS. This metric is then advertised to the other neighboring AS through the control plane using BGP extensions. BGP ensures that the information is percolated to other AS. On the receipt of these PWR metrics by the AS at far-end of the Internet, the overall AS level topology can be constructed. Note that this view of the Internet is available with each of the routers without using any other complex discovery mechanism. Some sample link weights shown in Figure 2 are obtained by using such a mapping function on the ingress links.

3) Low-power path detection: The algorithm consists of two sub-algorithms: each one executed by the ASBRs and the Path Computation Elements (PCEs) in the network in their respective AS. PCEs have been proposed by the Internet Engineering Task Force for path computation activities. We can use the existing PCE architecture for our algorithm. The algorithms for the ASBRs and PCEs are given as Algorithm 1 and Algorithm 2, respectively.

In Algorithm 1, parallel process 1 (steps $4-10$) is used to exchange the PWR metric information. Parallel process 2 (steps $11-16$) handles the TE-LSPs. Algorithm 2 calculates the low-power path from the head-end to the tail-end and sends this path information to the head-end AS.

Algorithm 1 ASBR low-power path algorithm

Require: Weighted Topology Graph T=(AS, E, f)

```
1: Begin
2: /* As part of Interior Gateway Protocol-Traffic Engineering */
3: Trigger exchange of available bandwidth on bandwidth change, to the AS internal neighbors;
4: BEGIN PARALLEL PROCESS 1
5: while PWR metric changes do
6:     Assign the PWR metric to the Ingress links;
7:     Exchange the PWR metric with its external neighbors;
8:     Exchange the PWR metric with AS's (internal) ASBRs;
9: end while
10: END PARALLEL PROCESS 1
11: BEGIN PARALLEL PROCESS 2
12: while RSVP packets arrive do
13:     Send and Receive TE-LSP reservations in the explicit path;
14:     Update routing table with labels for TE-LSP;
15: end while
16: END PARALLEL PROCESS 2
17: End
```

Algorithm 2 PCE low-power path algorithm

Require: Weighted Topology Graph T=(AS, E, f)

Require: Source and Destination for inter-AS TE LSP with sufficient bandwidth

```
1: Begin
2: Calculate the shortest paths from the head-end to the tail-end using CSPF with PWR as a metric;
3: if no path available then
4:     Signal error;
5: end if
6: if path exists then
7:     Send explicit path to head-end to construct path;
8: end if
9: Continue passively listening to BGP updates to update T=(AS, E, f);
10: End
```

4) Path establishment: Using the PWR metric the low-power path is obtained by applying the CSPF algorithm. For example, in Figure 2, the path (A, B, D, G, H, X) is power efficient as the summation of the PWR metric in this path is minimum when compared with other paths in the graph topology. Of course, the routing choice will depend on the reservation of the bandwidth on this path. If available bandwidth exists to setup a TE-LSP, then the explicit path

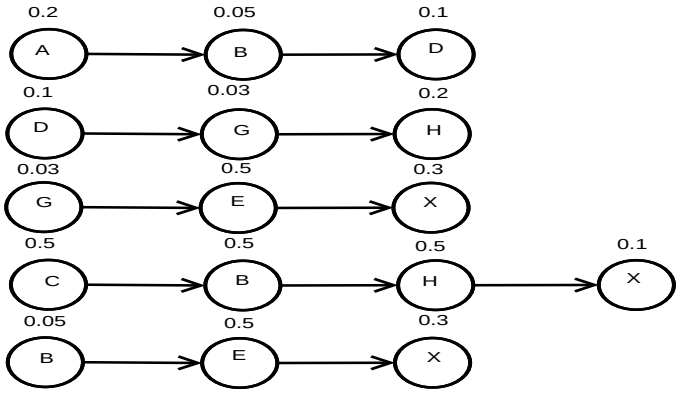


Fig. 1. Strands obtained from BGP updates, vertices A, B, C, D and G are the head-end AS; D, H and X are the tail-end AS. The vertex weights represent the PWR metric of an AS, and the link direction shows the next AS hop. ASBRs present the topology to the PCE.

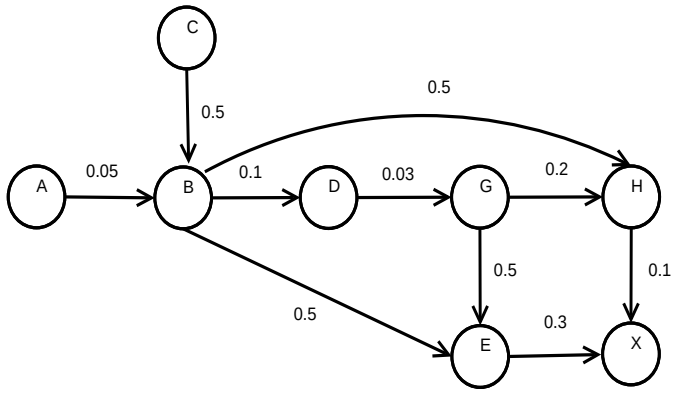


Fig. 2. Strands combined to get the Internet topology. The PWR metric is mapped to the ingress link of the ASBR. CSPF algorithm is run on this topology to detect the low-power path by the PCE.

is established. The Resource Reservation Protocol (RSVP) adheres to its usual operation and tries to setup a path. If bandwidth is not available in the low-power path thus calculated, then we fall back to the conventional shortest paths, provided there is available-bandwidth. The low-power path algorithm given as Algorithm 2 is executed by the PCE. Algorithm 1 prepares the topology and feeds it as input to the PCE as a weighted topology graph.

C. Time complexity

In protocol related algorithms two issues are of interest namely: communication and computation time complexity.

1) Communication time complexity: The communication time complexity involved in the algorithm includes

- monitoring the BGP packets to discover the topology based on AS-PATH-INFO attribute,
- exchanging PWR metric between the neighboring AS, and
- using inter-AS TE LSP to construct the path from the head-end to the tail-end AS.

Monitoring the BGP packets takes $O(1)$, a constant time. Exchanging PWR metric between the neighbors occurs in a distributed manner and hence takes $O(1)$, a constant time. The construction of the path takes $O(N)$ where N is the diameter of the network topology. Hence, the computational complexity is $O(N)$.

2) Computational time complexity: The algorithm for forming the topology from AS strands is dominated by the number of links. In the case of dense connectivity the computational time complexity is $O(|V| + |E|) \approx O(|E|)$, where $|E|$ is the number of edges and $|V|$ is the number of nodes or vertices. The computational time complexity of the low-power path algorithm is dominated by the Dijkstra's algorithm. In this case, instead of hops or any other metric we use the PWR metric in the Dijkstra's shortest path algorithm. Hence, the computational time complexity is bounded by Dijkstra's shortest path algorithm which is $O(|E| \log |V|)$.

The algorithm discussed above can be implemented "offline". The topology information could be extracted by passive monitoring, PWR metric information can be obtained using SNMP, and the low-power path can be calculated using a separate offline system. The paths can then be established by installing routing tables remotely. This offline implementation has certain drawbacks. We list the drawbacks and some possible solutions to overcome them.

1) Using the CSPF algorithm to calculate the route from source to destination could be time consuming for large networks. But the topology is dynamically updated and hence the computation of the shortest path can be triggered based on need.
2) The topology information obtained using the BGP strands might be incomplete. For a detailed discussion on completeness of Internet topology using BGP refer to [19], [20]. In addition to the BGP based algorithm, any other algorithm such as SNMP based Topology discovery could also be used to enhance connectivity as well as discover new nodes. But this increases the communication time complexity.
3) The PCEs usually use modified Dijkstra's shortest path algorithm and not a distributed algorithm. The algorithm can be speeded up using the graph-labeling method discussed in [1].
4) The algorithm uses RSVP and inter-AS TE LSP to establish the path which leads to communication overhead.

As we can see, too many information exchanges happen to implement the low-power path selection process. It would be of interest to see whether we could reduce the communication time complexity issues. Note that the computational time complexity is still bounded by the shortest path algorithm. In the next section, we explain the BGP path selection algorithm which overcome these issues.

III. BGP LOW-POWER PATH SELECTION

Before we study the proposed changes to the BGP based path selection algorithm, we will review the current algorithm

discussed in [16]. These algorithms are executed by the ASBRs and the core routers.

A. BGP path selection

In the BGP algorithm [15], each entity exchanges the best route to a given destination with other connected entities. Therefore, the BGP protocol is effectively a distributive method for generating routing information and there is no need to explicitly discover the topology. Of course this means that the information obtained from the neighboring entities must be reliable which is the case in the Internet. Such a distributed BGP algorithm exchanges prefixes and their next hops after going through the best path selection steps. Hence, there is a need to compare and choose the best route to add to the IP routing table which is used for routing the data packets. For this process to take effect, BGP uses about thirteen different rules to choose the path [16], [17]. We add the PWR metric-based low-power path selection as another rule.

The algorithm works as follows: BGP assigns the first valid path as the current best path based on the paths it received from the neighboring entities. BGP then compares the best path with the next path in the list, until BGP reaches the end of the list of valid paths. The rules that are used to determine the best path are given briefly in Algorithm 3.

There are some exception conditions in some of the steps; for details refer [16]. We now modify the BGP path selection algorithm functionality by including the low-power path PWR metric-based calculation. This involves adding Algorithm 1 as a subroutine to the BGP path selection criteria after line 4 and expanding line 5 of Algorithm 3, where we select the shortest path only if a low-power path is not available. The following conditions are considered in the PWR metric-based low-power path selection (see Algorithm 4).

1) If the PWR metric is not available for a link, we drop all paths using the link (steps $20-22$).
2) If the PWR metric-based detection is not enabled in even a single entity that uses BGP we do not execute this algorithm (steps $5-6$).
3) If there is only one AS PATH then we skip applying this algorithm (default action).
4) If there are multiple paths to the tail-end then we choose the one with the least sum of PWR metric to the tail-end (step 14).
5) If multiple path exists with the same PWR sum, then we choose all the paths and give it to the path selection algorithm (steps $13-18$).
6) If there are no PWR metric-based paths we fall back to the shortest path algorithm (default action).

The detailed changes are given as Algorithm 4. Algorithm 3 is now complete with the inclusion of the low-power path selection process using the PWR metric.

Note that this method involves changes to the BGP path selection algorithm and hence all the devices involved in exchanging BGP routes must implement this method. Therefore, this method cannot be implemented offline. We will refer to this as "online" implementation.

Algorithm 3 Abridged BGP algorithm

Require: Topology information related with BGP

1: Begin
2: Prefer the path with the highest WEIGHT a locally configured parameter for a router.
3: Prefer the path with the highest LOCAL_PREF value, a value configured for local preference.
4: Prefer the path that was locally originated via a network or aggregate BGP subcommand or through redistribution from an Interior Gateway Protocol.
5: Prefer the path with the shortest AS_PATH, the shortest path from source to destination.
6: Prefer the path with the lowest origin type (Exterior Gateway Protocol paths preferred over Interior Gateway Protocol paths).
7: Prefer the path with the lowest multi-exit discriminator (MED). This parameter is used when there are multiple paths to a destination.
8: Prefer external BGP over internal BGP paths.
9: **if** bestpath is selected **then**
10: go to MULTIPATH;
11: **end if**
12: Prefer the path with the lowest IGP metric to the BGP next hop.
13: MULTIPATH: Determine if multiple paths require installation in the routing table for BGP Multipath.
14: **if** best path selected **then**
15: exit with the best path.
16: **end if**
17: When both paths are external, prefer the path that was received first (the oldest path).
18: Prefer the route that comes from the BGP router with the lowest router ID.
19: If the originator or router ID is the same for multiple paths, prefer the path with the minimum cluster list length. The router ID is the highest IP address on the router, with preference given to loopback addresses.
20: Prefer the path that comes from the lowest neighbor address.
21: End

B. Time complexity

We now discuss the communication and computational complexity of the proposed algorithm.

1) Communication time complexity: Topology discovery is not needed in this algorithm as BGP exchanges best routes with the neighboring entities. This removes the need for using TE-LSPs to establish the path from the head-end to the tail-end AS as well. Of course, traffic engineering techniques can still be enforced. There is no additional communication overhead other than the addition of a new BGP attribute to the BGP protocol. Therefore, the total communication complexity is bounded by a constant, $O(1)$.

Algorithm 4 Modified BGP path selection algorithm

Require: BGP path selection algorithm

```
Begin
if ROUTER is configured with BGP then
    Execute Step 2, 3 and 4 of Algorithm 3
end if
if there is no PWR metric-based path selection then
    Goto Step 5 of Algorithm 3;
else
    if (there are multiple AS_PATHS) AND (PWR metrics) then
        Calculate the sum of PWRs in the paths.
    else
        Ignore paths that have no PWR metrics.
    end if
    if there exists multiple sum of PWRs as there is more than one path then
        Choose the AS_PATHS with the least PWR metric sum.
        if multiple least PWR metric sum are equal then
            Choose all the AS_PATHS;
            Goto Step 6 of Algorithm 3;
        end if
    else
        if there exist no PWR_SUM because of exclusion then
            Goto Step 5 of Algorithm 3;
        end if
    end if
end if
Goto Step 6 of Algorithm 3;
End
```

2) Computational time complexity: The computational time complexity of the BGP path selection algorithm is bounded by the calculation of the path with the smallest sum of PWR metric, if multiple paths exist. In the worst case, we might have to apply Dijkstra's shortest path algorithm with PWR metric on the topology learned by the ASBRs. Therefore, the computational time complexity is still bounded by that of Dijkstra's algorithm which is $O(|E| \log |V|)$, with $|E|$ and $|V|$ representing the number of edges and nodes, respectively. By using the proposed algorithm, we overcome the draw backs of the inter-AS TE-LSP based low-power path algorithm. Note that the path selection is done by the ASBR and there is also no need for the use of PCE in this method.

We conducted simulations using the offline PWR based method to study the possible power reduction.

IV. Simulations

The simulations involved creating various graph topologies for a given connectivity. For large values of vertices V GNU scientific library based simulation was performed [21]. We assumed a uniform link distribution between the nodes. Uniform distribution was used to ensure that there were minimal number of disconnected components under low connectivity. On this topology, PWR metric values based on uniform and exponential distribution were assigned to the links. The experiments were repeated for different set of values of distributions. We considered about 100 topologies for each connectivity ranging from 20% to 90%. Any graph topology that was disconnected was dropped from the study. We used the Dijkstra's algorithm for finding the low-power as well as shortest paths. The simulation parameters are given in Table I.

TABLE I
Simulation parameters and their values

Parameter	*Value*
Topology size	$100, 10000,$ 1million nodes
Connectivity	$25 - 95\%$, step size 5%
Low-power nodes	Uniform, Exponential ($\lambda = 0.25$)
Network types	100 topologies for each connectivity

Two important parameters were monitored: the increase in the number of hops and the comparative power reduction possible by opting for the low-power path algorithm. For each source-destination pair in the topology, we compared the power reduction obtained by using low-power paths with that of the conventional hop based shortest path metric. The PWR metric can vary dynamically over a time period which also means that the low-power paths can vary for a given connectivity. Therefore, we also monitored the power-reduction and hop variations for a connectivity of 70%. The graphs presented are for node size 100. For 70% connectivity we show 30 randomly chosen sample topology out of the 100 topologies that were used.

A. Uniform distribution of PWR metric

The graphs shown in Figure 3 for uniform distribution of PWR metric depicts the power reduction and hop increase for various connectivity size. We see that the minimum power reduction that can be achieve is around 10% and this increases almost linearly with connectivity. The hops can also increase by up to 50%. High values of power reduction were possible as there are equal number of links with high and low PWR metric under this distribution. It can be seen that the average number of hops increases with more connectivity. This is because more low-power links also increase under such circumstance and the algorithm prefers routing through such low-power links.

B. Exponential distribution of PWR metric

In this case, the topology had more low-power links. From Figure 4, it can be seen that the average power reduction as well as hop increase could be as high as 65%. This is possible as the network topology is biased towards low-power links. The hops also increase as the proposed method tries to use all the low-power links for establishing routing information. The simulations for the two distributions establish that the algorithm uses low-power paths to route data packets. It should be noted that the PWR metric uses the power consumption of

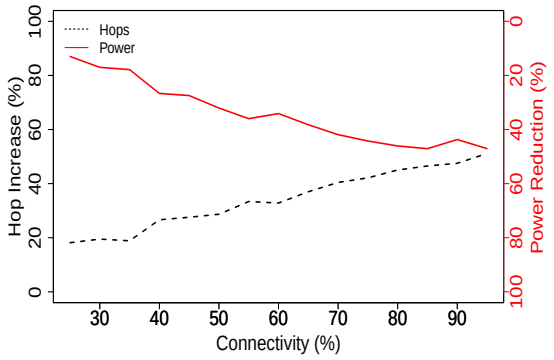


Fig. 3. Uniform distribution of power links

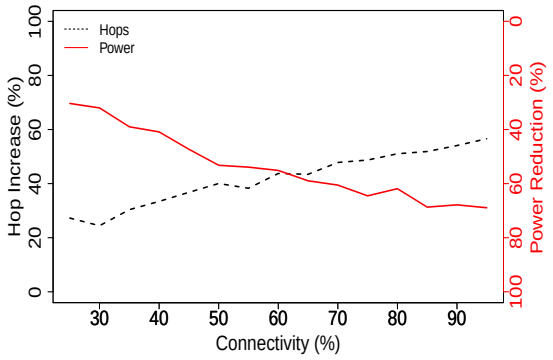


Fig. 4. Exponential distribution of power links

the AS. Even though the simulations can be considered very optimistic, typically AS consume Mega-Watts of power [22], [23]. Therefore, even a 1% reduction in power of an AS can result in significant savings for ISPs.

These results also suggest that after a particular value of connectivity size, increase in hops does not return much benefit with respect to power reduction. Therefore, it is interesting to study the behavior of the algorithm under a given connectivity value. Such a study will help to understand the dynamics of the network. Note that the PWR metric will also fluctuate over time and hence the paths can dynamically change.

C. Role of connectivity

We fixed a connectivity of 70% and studied the network for power reduction and hop increase using both the distributions (see Figures 5, 6). Results indicate that the graph topology plays a major role in power reduction. For both the distributions, the power reduction as well as increase in hops is bounded by a range which is dictated by the connectivity. The PWR metric will vary over a period of time. Therefore, the algorithm also reschedules the routing information based on the low-power values. Each trial can even be considered as the network topology at a time instant. The average power reduction remained quite high in both the cases.

We now discuss some implementation issues of the proposed online algorithm.

V. Implementation

In this section, we present notes on feasibility of implementation in a live network. We also briefly discuss implementation work based on OpenFlow [24] for offline implementation and Quagga [25] based online implementation.

A. Feasibility of implementation

First, the requested bandwidth should be available on the low-power path. This can be taken care using TE methods. Second, there is a reliable flooding process that gets triggered when updates about the change in PWR metric arise. We propose addition of some attributes with no change to the protocol implementation. There may be a time lag when the far ends of the Internet receive the attribute and the time it originated. This cannot be avoided as with other attributes and metrics. In MPLS-TE, when the TE metrics are modified, there is a reliable flooding process within an Interior Gateway Protocol (IGP). Such triggered updates apply to the PWR metric as well. The proposed PWR metric is advertised to the neighboring AS and the information percolated to all the AS, in a AS-PATH-POWER-METRIC attribute. This attribute is discussed in the Appendix. The frequency of the updates for this attribute should be fixed to avoid network flooding.

The AS-PATH-POWER-METRIC for each ASBR is calculated, and advertised as the PWR metric for the AS. This AS-PATH-POWER-METRIC is filled into an appropriate transitive non-discretionary attribute and inserted into a unique vector for a set of prefixes advertised from the AS. Such advertised prefixes may have originated from the AS or be the transit prefixes. The filled vector is sent to the ASBR of the neighboring AS, and later propagated to all the ASBRs. If the elements denoting AS in a vector of AS-PATH-INFO is not the same as the ones that need to be advertised in a AS-PATH-POWER-METRIC, then a suitable subset of AS-PATH-POWER-METRIC is identified and sent in the BGP updates. A vector of size 1 also can be employed if the AS in question is the only one for which PWR metric has changed in the originating AS.

The power consumed by each router may fluctuate over short time intervals. This can occur if the data packets are re-routed. In this case a low-power path might start consuming higher power and advertise a higher PWR metric. It is possible that the routes can flap due to PWR metric changes. In order to dampen these fluctuations, power can be measured when falling within suitable intervals as opposed to a discrete quantity. This method of power measurement reduces the frequency of triggered updates from the routers due to power change. This can sometimes affect the network performance. This situation must also be addressed while using PWR metric-

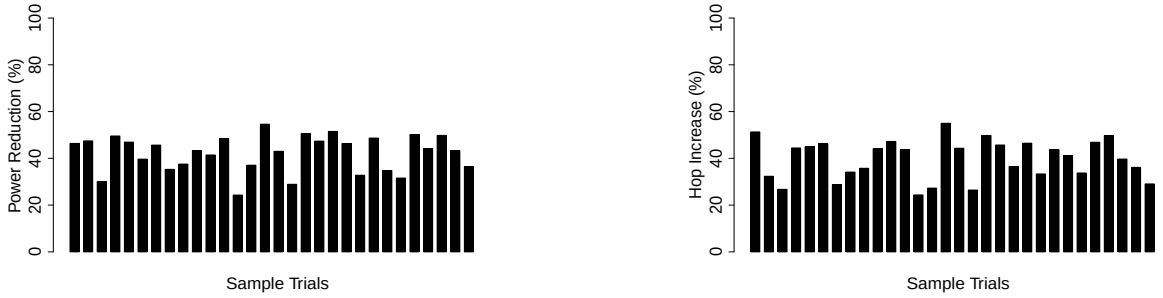


Fig. 5. Power reduction and hop increase for *uniform distribution* of power values with 70% connectivity.

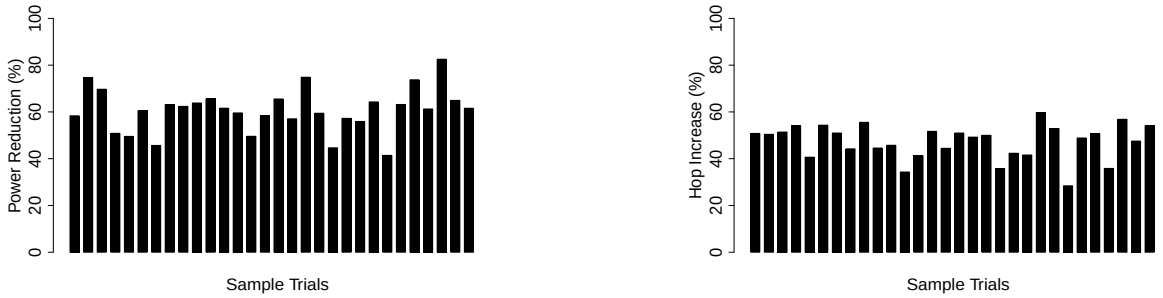


Fig. 6. Power reduction and hop increase for *exponential distribution* of power values with 70% connectivity.

based methods in the network.

Multiple ASBRs advertising differing PWR metric can lead to AS that have low PWR metric through an ingress link and not through other. Consider the case of multiple ASBRs that belong to the same AS, advertising differing PWR metrics. This could lead to power values that belong to different classes with intervening classes in between. These advertised PWR metrics could lead to one ASBR being preferred over the other thus taking a different path from head-end to tail- end. This also entails that there may be multiple paths to the AS through these different ASBRs. As an example, consider Figure 7 which shows a set of strands that derive a topology as in Figure 8. Here, D is reachable via two paths but the PWR metrics differ. This illustrates the case where the better metric wins out. The average power consumed would not have an effect but the bandwidth available on these ASBR egress links would definitely influence the path.

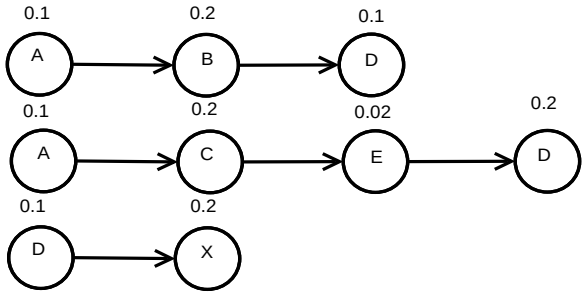


Fig. 7. Example of strands where more than one PWR metric is advertised by D.

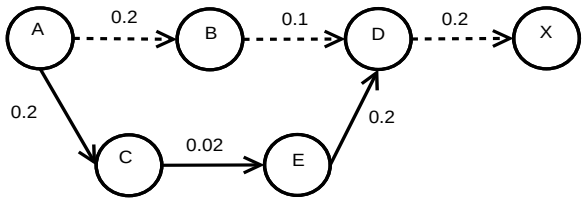


Fig. 8. Low-power path derived using the algorithm that uses low value ingress link but through the same AS.

B. OpenFlow implementation

Since we did not have access to a live network, we emulate the algorithms using a simple offline implementation based on OpenFlow [24]. OpenFlow was designed to run experimental protocols on the campus network for research purposes. Many of the vendor devices support OpenFlow as a part of their capability. The control flow part of the router/switch is handled by a OpenFlow controller while the data path still resides at the router. OpenFlow devices and the OpenFlow controller communicate with each other using OpenFlow protocol [24].

In our implementation, the router/switch are treated as Autonomous Systems. These devices present information about

their consumed-power to the centralized controller. To simplify the implementation, we assume that enough bandwidth is available at these routers. This is a realistic assumption as enough bandwidth is usually available in the core routers. The centralized controller determines the topology of the network based on the connectivity information obtained from the routers. Using this topology as well as the consumed-power information, the centralized controller updates the routing tables so that data packets traverse through low-power paths.

C. Quagga based implementation

As a part of online implementation strategy, Quagga based implementation is studied [25]. The BGP daemon is modified and routes are formulated based on the low-power path criteria. The power information is obtained through the use of an experimental MIB included in the Quagga based Linux routers. At a later stage we plan to incorporate the AS-POWER-PATH-METRIC discussed earlier.

VI. Discussion

In this section, we first compare the offline and the online approaches. We then outline Quality of Service (QoS) aspects that need to be considered to compensate for any increase in the number of hops when low-power paths are chosen.

A. Comparison of offline and online approaches

In the offline approach, TE-LSPs are needed to establish the path between the source and the destination. In contrast, in the online approach, separate TE based label switching can be completely avoided. The routing table is generated using the BGP algorithm itself and hence the overhead for establishing the paths is reduced considerably. To incorporate the online approach the BGP path selection algorithm has to be modified in all the core and edge routers implementing this scheme. The online approach is considerably faster than the offline approach with a trade off in the implementation complexity.

B. Latency in the network

Finding low-power paths in the graph topology might lead to an increase in the number of hops between a source and a destination. An increase in the number of hops could lead to an increase in queuing and propagation delay. Propagation delay is unavoidable as it depends on the transmission medium. Queuing delay is introduced rather naturally due to the store and forward design of Internet routers and also as a consequence of the design of the flow control methods implemented in transport protocols. Latency is a key QoS metric, and thus minimising end-to-end delay is an important network engineering task. To compound the problem, router buffers are currently sized based on an out-dated bandwidth-delay product rule which was intended to maintain full link utilisation. There are two options to reduce queuing delays: either reduce the buffer sizes dramatically [7], or make judicious use of feedback, from queues, to design better queue management policies [8]. A low-latency network thus improves QoS and could possibly enable the deployment of power saving methods which might require an increase in the number of end-to-end hops. Internet QoS is an area of active research among the communication networks community.

VII. Outlook

We propose a method, which employs a collaborative approach between AS, to reduce power consumption by using the Consumed-Power to Available-Bandwidth (PWR) metric.

A. Contributions

In our previous work [1], the AS topology was represented as a graph using the strands obtained from the AS-PATH-INFO attribute of the BGP updates. The CSPF algorithm was run on this topology by using the PWR metric as an additional constraint. The PWR metric is advertised through the ingress links of the ASBRs associated with AS using BGP updates. Inter-AS Traffic Engineering Label Switched Paths were used to route the data packets from the head-end to the tail-end. As using CSPF can be time consuming a heuristic algorithm to derive the low-power paths using graph-labelling was proposed. The communication time complexity associated with information exchanges in this method is high.

In order to reduce this complexity, in this paper we proposed that the BGP path selection algorithm be used to determine the low-power consuming paths between AS using the PWR metric. To study the performance and viability of using PWR metric-based methods, we conducted simulations on various topologies with different PWR metric distributions. The distributions used were the uniform and exponential distributions, and the results were especially encouraging: there was a substantial gain in power reduction where the tradeoff was an increase in the number of hops. We also briefly discussed emulating these schemes with OpenFlow and Quagga based BGP. Given the current power consumed by the AS, reduction in power savings could be rather beneficial to the ISPs.

B. Avenues for future work

The methods proposed in this paper assume that the PWR metric information is reliable. An erroneous metric information can be a cause of concern. However, ISPs usually have Service Level Agreements (SLAs) for carrying traffic. One method is to link up each ISP with a power application level gateway to ensure that proper metrics are advertised. This could be mandated at least amongst the cooperating ISPs.

It would be of interest to study whether the conceptual methods used at inter-AS level can be employed to inter-Area based topology. It is also natural to extend the study for multicast traffic. It would certainly be interesting to perform an evaluation of the proposed methods on a range of topologies and PWR metric distributions. Our work focused on the core and access networks that use BGP as the routing protocol. A study on extending these methods to other access networks implementing wireless connections would be useful. We have not considered the role of different traffic distributions on

power consumption. A practical study could be conducted on a live AS topology.

It has recently been highlighted that queuing delay in the Internet is on the rise [26]. The proposed scheme for power reduction would lead to an increase in the number of hops. Thus significant queuing delays at each hop would negatively impact QoS if the number of hops, between source and destination, are increased. Given the potential benefits for power reduction it would be imperative to investigate the design of queue management schemes to ensure a low latency network. Some work in this direction has already been started [8].

ACKNOWLEDGMENT

Shankar Raman would like to acknowledge the support by BT Public Limited (UK) under the BT IITM PhD Fellowship award. Balaji Venkat and Gaurav Raina would like to acknowledge the UK EPSRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC. We thank Prof. Kamakoti for allowing the use the RISE lab facilities for our simulations. We appreciate the helpful suggestions by Fabrice Saffre and Hanno Hildmann on the presentation.

REFERENCES

[1] S. Raman, B. Venkat, and G. Raina, *Reducing power consumption using the Border Gateway Protocol*, Proc. of the Second International Conference on Smart Grids, Green communications and IT Energy-aware Technologies, Energy 2012, March 2012, pp. 83–89, ISBN: 978-1-61208-189-2.

[2] J. Baliga, K. Hinton, and R. S. Tucker, *Energy consumption of the Internet*, Proc. of joint International Conference on Optical Internet, June 2007, pp. 1–3, doi: 10.1109/COINACOFT.2007.4519173.

[3] A. P. Bianzino, C. Chaudet, D. Rossi, and J. Rougier, *A survey of green networking research*, IEEE Communications and Surveys Tutorials, preprint, 2011, pp. 1–18, doi: 10.1109/SURV.2011.113010.00106.

[4] J. Chabarek, J. Sommers, P. Bardford, C. Estan, D. Tsiang, and S. Wright, *Power awareness in network design and routing*, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457–465, doi: 10.1109/INFOCOM.2008.93.

[5] G. Appenzeller, *Sizing router buffers*, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.

[6] W. Lu and S. Sahni, *Low-power TCAMs for very large forwarding tables*, IEEE/ACM Transactions on Computer Networks, vol. 18, no. 3, June 2010, pp. 948–959, doi: 10.1109/TNET.2009.2034143.

[7] G. Raina, D. Towsley, and D. Wischik, *Part II: control theory for buffer sizing*, ACM SIGCOMM Computer Communications Review, vol. 35, no. 3, July 2005, pp. 79–82, doi: 10.1145/1070873.1070885.

[8] S. Raman, S. Jain, and G. Raina, *Feedback, transport layer protocols and buffer sizing*, Proc. of the Eleventh International Conference on Networks, February 2012, pp. 125–131, ISBN:978-1-61208-183-0.

[9] A. Cianfrani, V. Eramo, M. Listanti, and M. Polverini, *An OSPF enhancement for energy saving in IP networks*, Computer Communications Workshops, Proc. of the IEEE INFOCOM 2011, April 2011, pp. 325–330, doi: 10.1109/INFCOMW.2011.5928832.

[10] R. Bolla, R. Bruschi, A. Cianfrani, and M. Listani, *Enabling backbone networks to sleep*, IEEE Network, vol. 25, no. 2, March/April 2011, pp. 26–31, doi: 10.1109/MNET.2011.5730525.

[11] R. Bolla, R. Bruschi, F. Davoli, and F. Cucchietti, *Energy efficiency in the future Internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures*, IEEE Communications Surveys and Tutorials, vol. 13, no. 2, second quarter 2011, pp. 223–244, doi: 10.1109/SURV.2011.071410.00073.

[12] M. Xia, M. Tornatore, Y. Zhang, P. Chowdhury, C. Martel, and B. Mukherjee, *Greening the optical backbone network: A traffic engineering approach*, IEEE ICC Proceedings, May 2010, pp. 1–5, doi: 10.1109/ICC.2010.5502228.

[13] G. Y. Li et.al., *Energy-efficient wireless communications: tutorial, survey, and open issues*, IEEE Wireless Communications, vol. 18, no. 6, 2011, pp. 28–35, doi: 10.1109/MWC.2011.6108331.

[14] B. Venkat, A. V. Rajagopalan, and B. Bhikkaji, *Constructing disjoint and partially disjoint InterAS TE-LSPs*, USPTO Patent 7751318, Cisco Systems, 2010.

[15] Y. Rekhter and T. Li,A border gateway protocol 4 (BGP-4). *http://tools.ietf.org/html/rfc4271*. [Accessed: December 6, 2012].

[16] BGP path selection algorithm, *http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094431.shtml*, last accessed [December 6, 2012].

[17] BGP path selection algorithm, *http://www.juniper.net/techpubs/en_US/junos11.4/topics/reference/general/routing-ptotocols-address-representation.html*, last accessed [December 6, 2012].

[18] F. Blanquicet and K. Christensen, *Managing energy use in a network with a new SNMP power state MIB*, IEEE Conference on Local Computer Networks, October 2008, pp. 509–511, doi: 10.1109/LCN.2008.4664214.

[19] H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, *Towards capturing representative AS-level Internet topologies*, Computer Networks, vol. 44, April 2004, pp. 737–755, doi: 10.1016/j.comnet.2003.03.001.

[20] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, *The (in)completeness of the observed Internet AS-level structure*, IEEE/ACM transactions on Networks, vol. 18, no. 1, February 2010, pp. 109-122, doi: 10.1109/TNET.2009.2020798.

[21] M. Galassi et.al., *GNU Scientific Library Reference Manual, 3rd Edition, ISBN: 0954612078*, http://www.gnu.org/software/gsl/manual/html_node/, last accessed [December 6, 2012].

[22] K. Hinton, J. Baliga, M. Z. Feng, R. W. A. Ayre, and R. S. Tucker *Power consumption and energy efficiency in the internet*, IEEE Network, vol. 25, no. 2, March-April 2011, pp. 6–12, doi: 10.1109/MNET.2011.5730522.

[23] A. P. Bianzino, L. Chiaraviglio, M. Mellia, and J. L. Rougier, *GRiDA: Green distributed algorithm for energy-efficient IP backbone networks*, Computer Networks, vol. 56, no. 14, 2012, pp. 3219–3232, doi: 10.1016/j.comnet.2012.06.011.

[24] N. McKeown, *OpenFlow: enabling innovation in campus networks*, ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, 2008, pp. 69–74.

[25] O. Bonaventure, *Software tools for networking*, IEEE Network, vol. 18, no. 6, 2004, pp. 4–5.

[26] K. Nichols and V. Jacobson, *Controlling queue delay*, Communications of the ACM, vol. 55, no. 7, 2012, pp. 42–50, doi: doi.acm.org/10.1145/2209249.2209264.

APPENDIX

The proposed AS-POWER-PATH-METRIC attribute is shown in Figure 9. Since the updates can be triggered quite frequently, sequence numbers are needed. The rest of the fields are needed to exchange the PWR information and are self-explanatory.

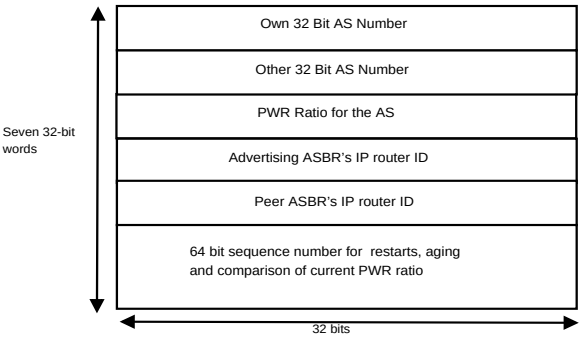


Fig. 9. AS-PATH-POWER-METRIC PDU

www.iariajournals.org

International Journal On Advances in Intelligent Systems

ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, ENERGY, COLLA, IMMM, INTELLI, SMART, DATA ANALYTICS

issn: 1942-2679

International Journal On Advances in Internet Technology

ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING, MOBILITY, WEB

issn: 1942-2652

International Journal On Advances in Life Sciences

eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO, SOTICS, GLOBAL HEALTH

issn: 1942-2660

International Journal On Advances in Networks and Services

ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION, VEHICULAR, INNOV

issn: 1942-2644

International Journal On Advances in Security

ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS

issn: 1942-2636

International Journal On Advances in Software

ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, IMMM, MOBILITY, VEHICULAR, DATA ANALYTICS

issn: 1942-2628

International Journal On Advances in Systems and Measurements

ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL, INFOCOMP

issn: 1942-261x

International Journal On Advances in Telecommunications

AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA, COCORA, PESARO, INNOV

issn: 1942-2601