CTRQ 2011

The Fourth International Conference on Communication Theory, Reliability, and Quality of Service

April 17-22, 2011

Budapest, Hungary

CTRQ 2011 Editors

Eugen Borcoci, Politehnica University of Bucharest, Romania

Pascal Lorenz, University of Haute Alsace, France

Michel Diaz, LAAS, France

CTRQ 2011

Foreword

The Fourth International Conference on Communication Theory, Reliability, and Quality of Service [CTRQ 2011], held between April 17 and 22 in Budapest, Hungary, continued a series of events focusing on the achievements on communication theory with respect to reliability and quality of service. The conference also brought onto the stage the most recent results in theory and practice on improving network and system reliability, as well as new mechanisms related to quality of service tuned to user profiles.

The processing and transmission speed and increasing memory capacity might be a satisfactory solution on the resources needed to deliver ubiquitous services, under guaranteed reliability and satisfying the desired quality of service. Successful deployment of communication mechanisms guarantees a decent network stability and offers a reasonable control on the quality of service expected by the end users. Recent advances in communication speed, hybrid wired/wireless, network resiliency, delay-tolerant networks and protocols, signal processing and so forth prompted the revision of some aspects of the fundamentals in communication theory. Network and system reliability and quality of service are those that mainly affect the maintenance procedures, on the one hand, and the user satisfaction on service delivery, on the other hand. Reliability assurance and guaranteed quality of services require particular mechanisms that deal with dynamics of system and network changes, as well as with changes in user profiles. The advent of content distribution, IPTV, video-on-demand and other similar services accelerate the demand for reliability and quality of service.

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

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

We hope that CTRQ 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the areas of communication theory, reliability and quality of service.

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

CTRQ 2011 Chairs:

Eugen Borcoci, Politehnica University of Bucharest, Romania
Alejandro Canovas Solbes, Polytechnic University of Valencia, Spain
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION - Zamudio, Spain

CTRQ 2011

Committee

CTRQ Steering Committee

Eugen Borcoci, Politehnica University of Bucharest, Romania
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Pascal Lorenz, University of Haute Alsace, France
Zary Segall, University of Maryland, USA
Michel Diaz, LAAS, France

CTRQ Advisory Chairs

Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION - Zamudio, Spain
Bogdan Ghita, University of Plymouth, UK
Raj Jain, Washington University in St. Louis, USA

CTRQ Industry/Research Chairs

Xuewen Gong, Huawei Technologies, China
Maria Striki, Telcordia Tehnologies Inc., USA

CTRQ Publicity Chair

Alejandro Canovas Solbes, Polytechnic University of Valencia, Spain

CTRQ 2011 Technical Program Committee

Paulo Alexandre Neves, Instituto de Telecomunicações/University of Beira Interior/Polytechnic Institute of Castelo Branco, Portugal
Mugurel Ionut Andreica, Politehnica University of Bucharest, Romania
Shingo Ata, Osaka City University, Japan
Himzo Bajric, Joint Stock Company BH Telecom - Sarajevo, Bosnia and Herzegovina
Jasmina Barakovic, BH Telecom, Joint Stock Company - Sarajevo, Bosnia and Herzegovina
Eugen Borcoci, Politehnica University of Bucharest, Romania
Julio César García Alvarez, Universidad Nacional de Colombia Sede Manizales, Colombia
Liliana Cucu-Grosjean, INRIA Nancy-Grand Est, France
Javier Del Ser Lorente, TECNALIA RESEARCH & INNOVATION, Spain
Philippe Devienne, CNRS, LIFL, Université de Lille I, France
Michel Diaz, LAAS, France
Manfred Droste, Universität Leipzig, Germany
Andrei Alexandru Enescu, Politehnic University of Bucharest, Romania
Andras Farago, The University of Texas at Dallas - Richardson, USA
Monique P. Fargues, Naval Postgraduate School - Monterey, USA
Markus Fidler, Leibniz Universität - Hannover, Germany
Bogdan Ghita, University of Plymouth, UK
Marc Gilg, Université de Haute Alsace, France
Xuewen Gong, Huawei Technologies, China
Alois Gosier, TU-Wien, Austria
Sascha Grau, Technische Universität Ilmenau, Germany

Amer Hassan, Microsoft Corporation, USA
Bjarne J. Helvik, NTNU, Norway
Benoît Hilt, IUT de Colmar, France
Robert C. H. Hsu, Chung Hua University, Taiwan
Mohsen Jahanshahi, Islamic Azad University - Tehran, Iran
Raj Jain, Washington University in St. Louis, USA
Brigitte Jaumard, Concordia University, Canada
Archana Kumar, Delhi Institute of Technology & Management - Haryana, India
Pascal Lorenz, University of Haute Alsace, France
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Adrian Matei, University POLITEHNICA of Bucharest / Orange Romania, Romania
Rick McGeer, HP Labs - Palo Alto, USA
Amalia N. Miliou, Aristotle University of Thessaloniki, Greece
Miklos Molnar, University Montpellier 2, France
Dritan Nace, Université de Technologie de Compiègne, France
Sarmistha Neogy, Jadavpur University, India
Petros Nicopolitidis, Aristotle University of Thessaloniki, Greece
Shahram Nourizadeh, LORIA-INPL, Nancy, France
Serban Obreja, University Politehnica of Bucharest, Romania
Jun Peng, University of Texas - Pan American - Edinburg, USA
Gianluca Reali, Università degli Studi di Perugia, Italy
Karim Mohammed Rezaul, Glyndwr University - Wrexham, UK
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Zahra Zohoor Saadat, Islamic Azad University-Pardis Branch, Tehran, Iran
Sébastien Salva, IUT d'Aubière - Cézeaux, France
Iraj Saniee, Bell Labs, Alcatel-Lucent - Murray Hill, USA
Panagiotis Sarigiannidis, University of Western Macedonia, Greece
Zary Segall, University of Maryland, USA
Dimitrios Serpanos, ISI/RC Athena & University of Patras, Greece
Vasco Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Joel Sommers, Colgate University, USA
Maria Striki, Telcordia Tehnologies Inc., USA
Tae Eung Sung, Cornell University - Ithaca, USA
Earl Swartzlander, University of Texas at Austin, USA
Mohammad Tehranipoor, University of Connecticut, USA
Pierre Tiako, Langston University - Oklahoma, USA
Kishor Trivedi, Duke University - Durham, USA
Elena Troubitsyna, Åbo Akademi University, Norway
Dimitrios D. Vergados, University of Piraeus, France Greece
Seppo Virtanen, University of Turku, Finland
Krzysztof Walkowiak Wroclaw University of Technology, Poland
Bernd E. Wolfinger, University of Hamburg, Germany
Yufeng Xin, University of North Carolina-Chapel Hill, USA
Abdulrahman Yarali, Murray State University, USA
Nicola Zannone, Eindhoven University of Technology, The Netherlands
Guangyan Zhao, Beijing University Aeronautics and Astronautics, China
Taieb Znati, University of Pittsburgh, USA

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Practical Performance Measurements and Analysis of IEEE 802.16 Networks

Jarmo Siltanen
JAMK University of Applied Sciences
Jyväskylä, FINLAND
jarmo.siltanen@jamk.fi

Kari Luostarinen
Metso Paper Inc.
Jyväskylä, FINLAND
kari.luostarinen@metso.com

Timo Hämäläinen
University of Jyväskylä
Jyväskylä, FINLAND
timo.t.hamalainen@jyu.fi

Abstract— **IEEE 802.16 standard defines the wireless broadband technology called WiMAX. When compared to other wireless technologies, it introduces many interesting advantages at physical (PHY) and media access control (MAC) layers. The WiMAX technology based on air interface standard 802.16 wireless metropolitan area network (MAN) is configured in the same way as a traditional cellular network with base stations using point to multipoint architecture to drive a service over a radius up to several kilometres. The range and the non line of sight (NLOS) ability of WiMAX make the system very attractive for users, but there will be slightly higher bit error ratio (BER) at low signal to noise ratio (SNR). WiMAX networks incorporate quality of service (QoS) mechanisms at the MAC layer. The problem of assuring QoS is basically that of how to allocate available resources among users in order to meet the QoS criteria such as delay, jitter and throughput requirements and how to achieve the optimal usage of resources to maximize throughput and to minimize power consumption while ensuring system scalability. In this paper, we make practical measurements of the WiMAX- system's QoS performance behaviour of the unsolicited grant service (UGS) and real time polling (RTP) classes.**

Keywords- IEEE802.16; WiMAX; Performance; Analysis

I. INTRODUCTION

IEEE 802.16, called also WiMAX, is a basic standard for the wireless broadband access network that can support a fast speed wireless access to different subscribers [1, 2].

The main benefit of WiMAX when compared to other wireless access network technologies like IEEE 802.11 are the longer range and more intelligent QoS support at the Medium Access Control (MAC) level [3]. There are several different types of applications and services, which can be used in the 802.16 networks and the MAC layer is designed to support this collaboration. An important feature of WiMAX is also that it is connection oriented and this means that an SS has to register to the base station before it can start to communicate with it. During the registration process, an SS asks the initial QoS requirements with the BS. These demands can be changed later if needed, and new connections can also be set up when needed. For providing the QoS guarantees in the WiMAX network the BS uses scheduling for both the uplink and the downlink channels. For that purpose there is an algorithm in the BS, which translates the QoS requirements of SSs into the appropriate number of slots [6-9].

According to the latest WiMAX Forum statistics, IEEE 802.16-based networks were deployed in 149 countries and regions, including 117 in Africa, 117 in Latin America, 109 in Asia-Pacific region, 86 in Eastern Europe, 76 in Western Europe, 53 in North America and 29 in the Middle East [13]. Global coverage using WiMAX service back in early 2009 had reached 430 million, to the end of 2010 more than 621 million WiMAX service coverage of population, according to forum latest estimates, 2011 will be expected to cover more than one billion users access to next-generation WiMAX networks.

The aim of this research is to compare the results received in WiMAX laboratory environments and in conjunction with laboratory exercise tests as well as the theoretical results from the mathematical modeling with the results received in authentic urban environments.

II. IEEE 802.16

A. General

The idea behind providing QoS in WiMAX lies on connection-oriented MAC architecture, service flow management and scheduling. Every time a SS and a BS need to communicate with each other, a unidirectional logical link is established. Each connection is mapped to a WiMAX service flow which works as a transport service for packets i UL or DL direction. The service flow defines QoS traffic parameters to be used for a connection. This include, for instance, traffic priority, maximum latency, tolerated jitter and maximum sustained traffic rate ensuring a specific level of service. Each service flow is identified by a 32-bit service flow identifier (SFID) [2].

Moreover, using adaptive modulation helps to utilize the bandwidth efficiently [4, 5]. WiMAX contains five QoS classes for the needs of various types of traffic: UGS , RTP, NRTP (non-real-time polling), ENRTP (enhanced non-real-time polling) and BE (best effort).

 ISBN: 978-1-61208-126-7

B. Different Traffic Types and Applications

The following scheduling types are applied to service flows, affecting the usage of UL bandwidth request opportunities. UGS - the Unsolicited Grant Service grants fixed-size UL allocations for an application minimizing the need for bandwidth requests, thus eliminating overhead and latency. This allows usage of real-time applications like VoIP. rtPS - the real-time Polling Service offers support for real-time UL transport with variable-size data packets. This makes it useful for video transmission. The rtPS optimizes data transport efficiency with a cost of slight overhead. ertPS - similarly to the UGS, in the extended real-time Polling Service the BS provides unsolicited unicast grants for the SS. Unlike the UGS, the ertPS allows dynamic-size UL allocations which makes it good for real-time voice and video applications. nrtPS - using the non-real-time Polling Service offers user unicast BW request opportunities on a regular basis, making it applicable even when the network resources are limited due to congestion. BE - the Best Effort service type allows SSs to use contention request opportunities. It is designed to be used when there is no minimum requirements for the connection. This service type suits for Web browsing. For each subscriber in the WiMAX network a QoS profile is defined and stored in the AAA server. This forms a basis for connection-oriented service, that is, associations between connections and service flows. The QoS profile stores allowable number of service flows, their scheduling types and values for other QoS parameters. [1][2].

Depending on the QoS profile and network properties the subscriber may use service flows provisioned via the network management system or dynamically created service flows. The provisioned service flows, including the initial service flow (ISF), can be created during the registration phase of the network entry after a successful authentication of the user. At this stage ASN-GW obtains SS´s QoS profile from AAA server which it uses to initialize service flow creation. Provisioned service flows can be activated or deactivated at any time when the SS is connected to the network (applicable to both static and dynamic service models). The dynamic service flow creation may be requested/initiated by the SS or by the network whenever a new connection is needed. DSA, DSC or DSD message exchange is used for service flow creation, modification or deletion, respectively [2].

III. PHYSICAL AND MAC LAYER QOS PROVISIONING

A. Physical layer basic concepts for QoS provisioning

Before transmission to the wireless link at the sender side, or right after reception at the receiver side, packets go through the IEEE 802.16 PHY layer. It performs operations, such as channel coding and interleaving, before passing on the packet. An adaptive physical layer is required to optimize the usage of resources, to accommodate user requirements and services and to maximize the spectral efficiency.

An adaptive modulation enables dynamic bandwidth allocation to match the current channel conditions. Modulation and coding scheme can be changed for each burst separately. Different modulation and coding schemes offer either robust or efficient network access, thus offering stable QoS in varying conditions. Three modulation schemes are are 64-QAM (quadrature amplitude modulation), 16-QAM, and QPSK (quadrature phase shift keying). The 64QAM offers highest bandwidth whereas the QPSK modulation is the most robust, therefore offering highest distance from the serving station. Also BPSK modulation can be used but it is not mandatory for UL or DL connections. In addition to different modulations there are a few coding rates used also to provide flexible networking. Depending on the Carrier-to-Noise Ratio (CNR) a coding rate of 1/2, 3/4 or 5/6 can be used.

The most commonly used technique for error correction is called forward error correction (FEC), which is capable of detecting and correcting some errors upon reception. This technique can reduce latency by cutting down the retransmissions, but the FEC requires more bits. The addition of FEC to every transmitted block reduces the efficiency of the channel and could increase the delay of good protocol data units (PDUs), on the other MAC-level ARQ can increase the delays in channel with high error rate. Many systems support hybrid techniques where a combination of FEC and ARQ parameters can be adjusted to allow the service requirements to be met under a variety of conditions [9].

WiMAX supports time division duplexing (TDD), full-duplex frequency division duplexing (F-FDD) and half-duplex frequency division duplexing (H-FDD). The usage of different duplexing modes affect the data transmission convention. In WiMAX user data is transmitted inside frames that consist of uplink and downlink subframes separated from each other by a TTG guard interval in TDD or different frequencies in FDD. Both DL and UL frames store user information into bursts which applies for both TDD and FDD frame. Depending on the duplexing mode being OFDMA/OFDM TDD, F-FDD or H-FDD, the frame structure changes. In addition to these, 802.16j defines a new frame format for communication between a multihop relay base station (MR-BS) and a RS [3].

In TDD the DL subframe and the UL subframe are transmitted in the same channel consecutively which is useful when working with limited bandwidth resources. In addition, TDD´s dynamically adjustable DL/UL ratio allows it to be used with both symmetric and asymmetric traffic. In FDD the DL and UL subframes are transmitted in separate concurrent channels using different frequencies. The difference between F-FDD and H-FDD is that in F-FDD a user is able to transmit and receive at the same time while in H-FDD user can either transmit or receive at given time but not do both simultaneously [8].

TDD is mostly used where WiMAX network is deployed. Main reasons for the popularity of TDD are flexibility, cost efficiency and high spectral efficiency [5].

 ISBN: 978-1-61208-126-7

Taking a closer look into a TDD frame reveals how user information is divided within a frame. The smallest time-frequency divided element in WiMAX is called a slot, which constitutes one subchannel and one, two or three OFDM/OFDMA symbols. All the data regions in a frame are composed from sequential slots.

The uplink subframe is the TDMA portion which may be used by one or more SSs to transmit information to the BS. Unlike the downlink, the UL-MAP grants bandwidth to specific SSs. The SSs transmit in their assigned allocation using the burst profile specified by the uplink interval usage code (UIUC) in the UL-MAP entry granting them bandwidth. The uplink subframe may also contain contention-based allocations. The number of contention slots per UL subframe is determined by the BS [7]. The BS assigns burst profiles for a specific connection or subscriber based on variety of constraints, including the QoS and channel conditions.

B. MAC layer concepts for QoS provisioning

Since the QoS requirements vary a lot in different network situations, WiMAX has many handling and transporting mechanisms to handle this. The IEEE 802.16e MAC uses a variable length PDU and multiple PDUs can be put into a single burst to save physical overhead. The MAC has a self-correcting bandwidth request/grant method that minimizes the overhead and delay of acknowledgements.

The MAC layer has three sublayers, which defines the access mechanisms and packet formats. These sublayers are service specific convergence sublayer (CS), MAC common part sublayer (MAC CPS) and MAC privacy sublayer (MAC PS). The CPS layer mainly interfaces with higher layer protocols, such as IPv4, IPv6 or ATM. The PS handles authentication and data encryption issues.

The IEEE 802.16 has two service-specific convergence sublayers that are used to map services to MAC connections. The first one is the ATM Convergence sublayer and the second is a Packet Convergence sublayer. It is used for IP, Ethernet, and virtual local area network environments. The goal of Packet Convergence sublayer is to classify SDUs and put them at the proper MAC connection. This ensures that QoS requirements are met and the bandwidth allocation takes place. The QoS requirements have the following features [6]:

a) A configuration and registration function for preconfiguring SS-based QoS service flows and traffic parameters.
b) A signaling function for dynamically establishing QoS-enabled service flows and traffic parameters.
c) Utilization of MAC scheduling and QoS traffic parameters for uplink service flows.
d) Utilization of QoS traffic parameters for downlink service flows.
e) Grouping of service flow properties into named Service Classes, so upper-layer entities and external applications (at both the MS and BS) may request service flows with desired QoS parameters in a globally consistent way. To meet all these requirements, the standard has three main methods: service flow QoS scheduling, dynamic service establishment and two-phase activation model.

IV. MEASUREMENTS IN URBAN ENVIRONMENTS

The measurement compared the behavior of the UGS and RTP classes upon overloading the link with 40 Mbps upstream BE traffic. The best effort traffic was generated with a Fluke Optiview analyzer over a WiMAX link. Simultaneously, a JDSU MTS-6000A analyzer was used to generate 70 VoIP telephone calls (G.729 codec) within the measured QoS class to a SmartClass Ethernet remote end. This number of phone calls with the above-mentioned codec amounts to a traffic flow of approx. 2 Mbps, and it was confirmed for both of the measured QoS classes (See Figure 1).

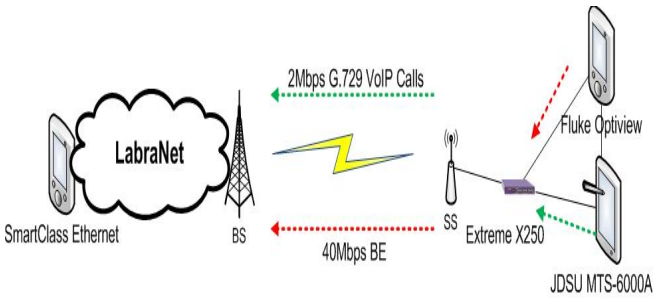


Figure 1. Measurement environment.

The MTS-6000A analyzer was used to read the delays, delay variations and throughput of the QoS class under observation. The measurement was conducted with adaptive and forced 16QAM3/4 modulation.

Measurement 1

The goal of the measurement was to receive results concerning the effect of modulation and channel width on data transfer speed in an urban environment. The aim was to compare the results with the theoretical results of mathematical modeling presented in [10].

The measurements were carried out with the base station installed on the roof of building 1 in Figure 3 and the subscriber station installed on the surrounding environments of Jyväskylä on building 2 in Figure 3 with LabraNet network active devices. Network topology was as in Figure 2. JDSU MTS-6000A traffic generator connected to c3750 switch was used to generate test traffic and the traffic was routed to front end of SmartClass Ethernet connected to a customer device. JDSU MTS-6000A traffic generator was controlled with VNC application from a NetSpan server connected to the control network.

 ISBN: 978-1-61208-126-7

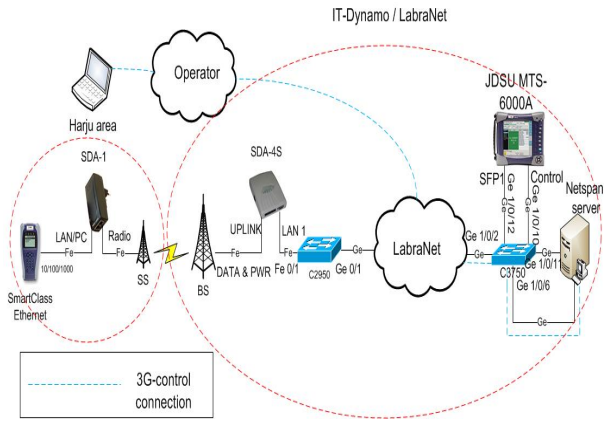


Figure 2. Network topology used in the measurements.

Table 1 shows the most important settings for the WiMAX devices.

Settings of the test environment	
Frequency	5.8 GHz
Channeling	FDD
Transmitting power	17.00dBm / 22.00 dBm
Frame duration	5 ms
Cyclic Prefix (Ratio G)	1/16
Distance of SS and BS devices	1 km
Test environment	City

Table 1 Test environment settings

Figure 3 shows the location and direction of the BS and SS device.

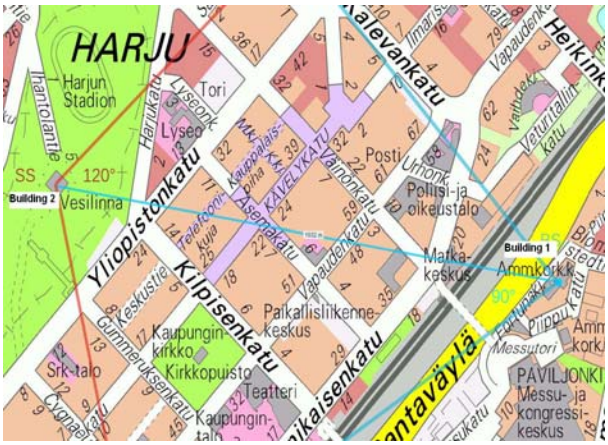


Figure 3. Location and direction of the BS and SS device.

Table 2 shows the main results of the measurement.

Modulation	Channel width	Data Transmission speed
16QAM 3/4	10 MHz	18.7 Mbps
16QAM 3/4	5 MHz	7.6 Mbps
16QAM 3/4	2.5MHz	2.2 Mbps
16QAM 1/2	10 MHz	12.5 Mbps
16QAM 1/2	5 MHz	5.1 Mbps
16QAM 1/2	2.5MHz	1.4 Mbps
QPSK 3/4	10 MHz	9.3 Mbps
QPSK 3/4	5 MHz	3.8 Mbps
QPSK 3/4	2.5MHz	1 Mbps
QPSK 1/2	10 MHz	6.2 Mbps
QPSK 1/2	5 MHz	2.5 Mbps
QPSK 1/2	2.5MHz	0.7 Mbps
BPSK 1/2	10 MHz	3.1 Mbps
BPSK 1/2	5 MHz	1.2 Mbps
BPSK 1/2	2.5MHz	0.3 Mbps

Table 2 Measurement results

The theoretical results of mathematical modeling (see Table 3) presented in [10] were compared with the actual measurement results shown in Table 2. A significant disparity was found between the theoretical and measured results in terms of the effect of channel width to data transfer speed per MHz.

Modulation	Channel width	Data transmission speed
16QAM 3/4	14 MHz	34.91 Mbps
16QAM 3/4	7 MHz	17.46 Mbps
16QAM 3/4	3.5 MHz	8.73 Mbps
16QAM 3/4	1.75 MHz	4.36 Mbps
16QAM 1/2	14 MHz	23.28 Mbps
16QAM 1/2	7 MHz	11.64 Mbps
16QAM 1/2	3.5 MHz	5.82 Mbps
16QAM 1/2	1.75 MHz	2.91 Mbps
QPSK 3/4	14 MHz	17.46 Mbps
QPSK 3/4	7 MHz	8.73 Mbps
QPSK 3/4	3.5 MHz	4.36 Mbps
QPSK 3/4	1.75 MHz	2.18 Mbps
QPSK 1/2	14 MHz	11.64 Mbps
QPSK 1/2	7 MHz	5.82 Mbps
QPSK 1/2	3.5 MHz	2.91 Mbps
QPSK 1/2	1.75 MHz	1.45 Mbps
BPSK 1/2	14 MHz	5.82 Mbps

 ISBN: 978-1-61208-126-7

BPSK 1/2	7 MHz	2.91 Mbps
BPSK 1/2	3.5 MHz	1.45 Mbps
BPSK 1/2	1.75 MHz	0.73 Mbps

Table 3 The theoretical results of mathematical modeling

In the measurements, the data transfer speed per MHz was highest at a channel width of 10 MHz. The speed was more than two times lower at a channel width of 2.5 MHz when measured with every available modulation. In the results of the mathematical modeling, the data transfer speed per MHz remained constant when the channel width was changed. On the other hand, the effect of modulation on data transfer speeds was found to be nearly identical between the theoretical and measured results. Table 4 shows the most important parameters of the mathematical modeling.

Parameters of mathematical modeling	
Channeling	FDD
Cyclic exponent	3
Frame duration	20 ms

Table 4 Parameters of mathematical modeling [10].

1) Measurement 2

The aim of the measurement was to compare the measurement results to the results of measurements conducted in a short-distance LOS environment in a laboratory. The equipment and topology used in the measurement are shown in Figure 2. Table 1 shows the most important settings for the WiMAX devices.

Table 5 lists the data transfer speeds achieved with the 64QAM, 16QAM, QPSK and BPSK modulations.

Modulation	Coding level	Bit rate (Mbps)
64QAM	¾	28.0
16QAM	¾	18.7
QPSK	¾	9.3
BPSK	½	3.1

Table 5 Measurement results for modulations

With 64QAM modulation, the results received in the urban environment showed only negligible disparity when compared to the measurement results obtained in the laboratory: a 28 Mbps downstream speed, as limited by the traffic profile, was achieved in both environments. With lower modulations, the laboratory speeds were 13-14% higher. The laboratory results are shown in Table 6.

Modulation	Coding level	Bit rate max (Mbps)	Average bit rate (Mbps)
64QAM	¾	28.0	26.0
16QAM	¾	21.2	21.2
QPSK	¾	10.6	10.6
BPSK	½	3.5	3.5

Table 6 Laboratory results

Table 8 shows the effect of channel width on the data transfer speed in an urban environment. The effect of channel width on the data transfer speed was nearly identical with the laboratory results listed in Table 7.

Channel width (MHz)	Modulation	Coding level	Bit rate (MHz)	Average bit rate (Mbps)
10	64QAM	¾	28.0	26,0
5	64QAM	¾	14.5	14,5
2.5	64QAM	¾	5,9	5,9

Table 7 The effect of channel width on the data transfer speed

The difference between the laboratory and urban environments when moving from a 2.5 MHz channel to a 5 MHz channel is only 0.1 Mbps. The transfer to 10 MHz cannot be taken into consideration as the 28 Mbps limit imposed by the traffic profile was reached in both environments.

Channel width (MHz)	Modulation	Coding level	Bit rate (MHz)
10	64QAM	¾	28.0
5	64QAM	¾	12.0
2.5	64QAM	¾	3.3

Table 8. Measurement results for channel widths

2) Measurement 3

The measurement tested the effect of direction, line-of-sight obstructions caused by trees, and transmission power on signal noise ratio in an LOS environment. The measurement examined downstream SNR values. The equipment and topology used in the measurement is as in measurement 1.

BS and SS device location and direction are shown in Figure 2. The SS device was directed to both sides of the BS

device at horizontal angles of 15°- 45° in different phases of the measurement. Initially, the measurements were conducted in three locations in the Harju (Figure 3) neighborhood where the line of sight to the BS device was obstructed in varying degrees by tree trunks and branches. The antenna element was aimed directly at the BS device through the obstructions.

The density of the line-of-sight obstruction caused by the tree trunks and branches did not have a significant effect on the results. Nearly identical values for downstream SNR were measured in all three locations. Only increasing the transmission power notably affected the downstream SNR values in a partial NLOS environment. We were able to increase the SNR value by directing the antenna element towards a tall building 15° to the southwest of the BS device, which means that the best SNR result was achieved by means of reflection. Table 9 shows the effect of line-of-sight obstructions and transmission power on the SNR value.

Coverage of the trees	Transmit power (dBm)	SNR
Large	17	24
Mean	17	24,2
Small	17	24,4
Large	22	27,6
Mean	22	28,0
Small	22	28,0

Table 9. The effect of LOS obstructions and transmission power on the SNR value

The subsequent measurements were conducted on the terrace of the building 2 in Figure 3 where a perfect LOS environment was achieved. When measuring from the terrace, the best SNR value was achieved when the antenna element was directed significantly to the side of the BS device's beam as this allowed reflections from the surrounding metal structures to strengthen the signal. Increasing the transmission power notably affected the SNR values in a perfect LOS environment, as well. After moving from a partial NLOS environment to a LOS environment, considerable differences were measured in the SNR values. Table 10 lists the SNR values measured in the LOS environment.

Transmission power (dBm)	SNR
17	28,5
22	33

Table 10. SNR values measured in the LOS environment

V. SUMMARY OF THE MEASUREMENT RESULTS

The results of the measurement were slightly marred by the behavior of the UGS class, for which many remedies were tried over several months. Airspan recommends that, for the UGS class, the employed VoIP codec's sample rate, which is 30 ms with the G.729 codec, should be used as the polling time. The recommendation proved to be sound, as the UGS traffic began running without background traffic after the polling time was changed even though no VoIP traffic was present in the class. However, changing the polling time did not resolve the entire problem, as a new obstacle presented itself. With the new 30 ms polling value, traffic ran even without background traffic, but if BE traffic, which overloaded the link, was added to the background, the UGS class delay and delay variation decreased while the frame loss throughput increased against all logic.

According to the manufacturer, the UGS class allows for reserving a certain amount of bandwidth within the framework of a particular polling time, as in the example below. Example: 64,000 bits per second (Max Sustained Rate and Min Reserved Rate) equals 8,000 bytes per second. Therefore, a bandwidth reservation of 800 bytes can be made for each polling cycle by setting the polling time to 100 milliseconds. With these settings, the packet size would be 800 bytes, from which overheads must also be subtracted. With the PING – lXXX command, the result can be confirmed at 601 bytes with headers, which is approx. 743 bytes over the link [11].

We tested the class' behavior with different polling values and found the limit to be 31 ms. After this, the class will not function without BE background traffic. However, overloading the link with BE traffic does not diminish the values of the UGS. In fact, it does the opposite: delays, delay variation and packet loss decrease while throughput increases.

VoIP service products should be created in such a way that the classifier directs all SIP (session initiation protocol) signaling that passes through port 5060 to the BE class, whose traffic priority is lower than that of VoIP traffic directed to the UGS class. However, in all of the devices concerned, signaling over the WiMAX link is assigned to predefined, locked control traffic classes whose traffic priority is 8 or 9, which is higher than the priority of any other traffic class [12].

Despite countless attempts, we were unable to get the UGS class to function properly independently. At this stage, the software on the base station device and subscriber stations was updated to its latest versions. In addition, the new Netspan server application was updated to its latest version. However, the updates did not have any apparent effect on the functioning of the classes. An error report describing the problem was sent to the manufacturer, but we never received a reply. The manufacturer's suggestion to direct the signaling to the BE flow and other VoIP traffic to the UGS flow would possibly enable the class to function without BE background traffic but it would not solve or explain the problems with the class' other behavior. In the absence of a reply from the manufacturer, it remains unclear as to why this important signaling should be incorporated into the BE flow, which contains no attributes pertaining to quality of service, using a lower priority than that of VoIP traffic in the UGS class. The manufacturer established a

 ISBN: 978-1-61208-126-7

remote connection to the Netspan server, but the problems were still not resolved and the solution suggestions were too short and ambiguous to be helpful.

In the earlier stages, when we were struggling with throughput problems, the manufacturer provided us with default configurations with which the devices were supposed to work as well as possible. However, the configurations proved to be unusable as, after they were implemented, the link functioned even worse than with our own configurations and the speeds dropped to a mere few hundred kilobits per second. After a more careful scrutiny of the manufacturer's website, we found that the 5.8 GHz licence-free frequency devices in our use had not been directly tested by Airspan. The test results represented similar devices using licensed frequencies.

Summary

In the course of the measurements, it was noted that the UGS class does not function without BE traffic in the background. Not even PING requests (packet internet groper) passed through the link in the UGS class alone without any background traffic. The UGS class traffic activated immediately after the introduction of 40 Mbps background traffic. Therefore, the activity of the class did not affect the measurements.

However, excluding the above-mentioned characteristic, the classes functioned properly. When using adaptive modulation for the measurements, the UGS class' delays remained approximately at a constant 22 milliseconds, whereas the delays of the RTP class rose significantly higher to over 90 milliseconds. Moreover, the RTP class presented a notably lower throughput – approx. 0.5 Mbps – compared to the nearly full 2 Mbps throughput of the UGS. This was directly reflected on the RTP class' frame loss, which was measured at 1.6.

Nevertheless, with 16QAM3/4 modulation, the differences between the classes reduced slightly. The delays of both classes dropped to below 20 milliseconds, but significant spikes occurred in the delay variation of the RTP class as opposed to the UGS class, which remained at a steady 2 milliseconds throughout the measurement. The same disparity was found regarding the frame loss, with the UGS class remaining near 0.1 and the RTP class presenting repeated spikes as high as 0.7.

VI. CONCLUSIONS AND FUTURE WORKS

Based on the results of the practical measurements, technology utilizing the licence-free frequency range could well be used to replace ADSL (asymmetric digital subscriber line) connections over copper cables in rural areas. At a distance of one kilometer, we reached a downstream data transfer speed of 28 Mbps, whereas with ADSL2+ technology, the data transfer speed a kilometer away from the DSLAM (digital subscriber line access multiplexer) is slightly over 20 Mbps, after which the speeds begin to drop dramatically as the distance increases. Currently, traffic at the 5.8 GHz frequency range is low in Finland. Therefore, the technology could also be used in urban areas, for instance, to connect two offices of a small or medium-sized business with a wireless WiMAX link with no monthly charge and with quality of service capability to ensure the uninterrupted flow of VoIP traffic, for example. After conducting various tests with the devices in a variety of environments, we found that the devices were best suited for use by businesses for Point-to-Point connections in an urban environment or for educational purposes due to their price and ease of use. Particularly in cities, businesses could replace costly ADSL connections that must be leased from operators with their own WiMAX devices. However, using the devices for the above-mentioned purposes is hindered by the abnormal behavior of the UGS class, which, without an update that remedies the problem, presents problems for placing VoIP calls and examining QoS classes.

Our future research will aim at optimizing the problem of cost-effective coverage area extension by using relays and consider novel resource management algorithms for multi-hop WiMAX networks.

ACKNOWLEDGEMENTS

We express our acknowledgements to Aku Aho and Juha Kuusenmäki from JAMK University of Applied Sciences Jyväskylä, Finland for contributing the practical measurements.

REFERENCES

[1] Air interface for fixed broadband wireless access systems. IEEE Standard 802.16, Jun 2004.

[2] Air interface for fixed broadband wireless access systems - amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands. IEEE Standard 802.16e, Dec 2005.

[3] K. Lu, Y. Qian, H.-H. Chen, and S. Fu. WiMAX neworks: from access to service platform. IEEE Network, 22(3):38–45, May/Jun 2008.

[4] Air interface for broadband wireless access systems. IEEE Standard 802.16 (Rev2 D5), Jun 2008.

[5] Air interface for broadband wireless access systems: Multihop relay specification. IEEE Standard 802.16j (D6), Jul 2008.

[6] IEEE 802.16m evaluation methodology document (EMD). IEEE 802.16 Broadband Wireless Access Group, Mar 2008.

[7] WiMAX Forum Mobile System Profile, Release 1.0 Approved Specification, Apr 2008. Revision 1.6.1.

[8] WiMAX Forum Network Architecture, Sep 2008. Release 1, Version 1.3.0.

[9] Ohrtman, F. 2005. WiMAX Handbook. Building 802.16 Wireless Networks

[10] P. Ritthisoonthorn, PhD dissertation: Designing Cost Effective Broadband Fixed Wireless Access Networks Through Engineering-Economic Analysis, Asian Institute of Technology, School of Engineering and Technology, Thailand, December 2009.

 ISBN: 978-1-61208-126-7

[11] Airspan Networks. 2010b. http://www.part-90.org/equipment/airspan/PIDASMAX36/user%20manual.pdf, 28th of March 2011.

[12] Airspan Networks. 2010d. ProST Hardware Installation Guide SR 6.5. http://www.part-90.org/equipment/airspan/PIDASMAX37/user%20manual.pdf, 28th of March 2011.

[13] WiMAX Forum, http://www.wimaxforum.org/, 28th of March 2011.

 ISBN: 978-1-61208-126-7

Performance Evaluation for DSRC Vehicular Safety Communication: A Semi-Markov Process Approach

Xiaoyan Yin[+], Xiaomin Ma[*], and Kishor S. Trivedi[+]
[+]Department of Electrical and Computer Engineering, Duke University, Durham, USA
[*]Department of Engineering and Physics, Oral Roberts University, Tulsa, USA
Emails: xy15@duke.edu, xma@oru.edu, kst@ee.duke.edu

***Abstract*— In this paper, an analytic model is proposed for the performance evaluation of vehicular safety related services in the dedicated short range communications (DSRC) system on highways. The generation and service of safety messages in each vehicle is modeled by an M/G/1 queue. A semi-Markov process (SMP) model is developed to capture contention and backoff behavior in IEEE 802.11 broadcast ad hoc networks. Furthermore, this SMP interacts with the M/G/1 queue through fixed point iteration. Based on the fixed-point solution, performance indices including transmission delay and packet delivery ratio (PDR) are derived. Hidden terminal problem is taken into account for the PDR computation. Analytic-numeric results are verified through extensive simulations under various network parameters. Compared with the existing models, the proposed model is more general and accurate.**

Keywords-analytic model; DSRC; performance evaluation; safety message; SMP model; VANET.

I. INTRODUCTION

Inter-Vehicle Communication (IVC), as a vital part of Intelligent Transportation System (ITS) [1], has been extensively researched in recent years. In the vehicular ad hoc network (VANET), the transportation safety is one of the most crucial features needed to be addressed. Safety applications usually demand direct vehicle-to-vehicle ad hoc communication due to highly dynamic network topology and strict delay requirements. Such direct safety communication will involve a broadcast service because safety information can be beneficial to all vehicles around a sender. Broadcasting safety messages is one of the fundamental services in dedicated short range communications (DSRC) [1], which is adopted by IEEE and ASTM.

The performance of vehicular safety communication in DSRC system has been studied in [2][3][4]. However, the evaluations are mainly based on simulations. Recently, a few analytic models based on discrete time Markov chain (DTMC) are developed in [5][6][7] to analyze the performance of the broadcast service incorporating the backoff counter process, hidden terminals and message generation interval. Nevertheless, these cited papers conduct performance assessments in a discrete time fashion by synchronizing system behavior to unit time slot, which will lead to some approximations in the results. In addition, according to 802.11 DSRC MAC layer protocol, a vehicle can directly transmit a packet without undergoing backoff process. Such phenomena has been ignored in the previous work [5][6][7], which will result in further approximations.

In this paper, we develop more accurate analytic models using a semi-Markov process (SMP) [8][9] interacting with an M/G/1 queue for the performance evaluation of the broadcast service in DSRC safety communication system. Fixed point iteration [10] is applied to derive the converged solution in the steady state. New approaches to calculate the transmission delay of safety related messages and PDR utilizing features of SMP models are also developed in this paper. The analytic results are verified by simulations.

This paper is organized as follows. Section II briefly describes the system behavior in 802.11 MAC layer protocol and assumptions in the system to produce a simplified model. Section III presents the analytic models and the fixed point iteration. Performance indices including mean delay and PDR are derived in Section IV. The analytic and simulation results are compared in Section V. Conclusions are presented in the last section.

II. SYSTEM DESCRIPTION

A. *Broadcast protocol*

In the 802.11 MAC layer protocol [11], distributed coordination function (DCF) is the primary medium access control technique for broadcast services. This section briefly explains the basic access mechanism of DCF for broadcast.

Each vehicle in the network can occasionally generate safety related packets and compete for the channel resource to transmit the packet. For a newly generated packet in a vehicle, the vehicle senses the channel activity before it starts to transmit the packet. If the channel is sensed idle for a time period of distributed inter-frame space (DIFS), the packet can be directly transmitted. Otherwise, the vehicle continues to monitor the channel until channel is detected to be idle for DIFS time period. Subsequently, according to the collision avoidance feature of the protocol, the vehicle generates an initial random backoff counter and goes through the backoff process before transmitting the packet. Moreover, a vehicle must go through the backoff process between two consecutive packet transmissions even if the channel is sensed idle for the duration of DIFS time for the second packet. Therefore, a packet can directly transmit without undergoing the backoff process only when the following two conditions are satisfied:

- The packet is generated when the queue is empty;
- The channel is sensed idle for DIFS time starting from the time instant that the packet is generated;

 ISBN: 978-1-61208-126-7

Regarding the backoff process for a packet transmission, the initial backoff counter is chosen randomly from a uniform density over the range (0, W_0-1). The backoff time counter is decreased by one if the channel is sensed idle for a time slot σ. The counter is *frozen* when channel is sensed busy and reactivated when the channel is sensed idle again for more than the DIFS duration. The packet is transmitted as soon as the backoff counter reaches zero.

In broadcast services, the transmitting vehicle does not receive any feedback from the receivers and will not retransmit a packet. The detailed descriptions for IEEE 802.11 standard can be found in [11].

B. System assumptions

Several assumptions are made in the broadcast system to produce a simplified yet a high fidelity model.

- The vehicular ad hoc network is considered to be one-dimensional. The number of vehicles in a lane is Poisson distributed with parameter β (vehicle density), i.e., the probability $P(i, l)$ of finding i vehicles in a lane of length of l is given by:

$$P(i,l) = \frac{(\beta l)^i}{i!} \cdot e^{-\beta l} \tag{1}$$

- All vehicles have the same transmission range as well as receiving range R.
- Each vehicle is assumed to generate packets as a Poisson stream with rate λ (in packets per second).
- Each vehicle has an infinite queue to store the packets at the MAC layer. Hence, each vehicle can be modeled as an M/G/1 queue, which has exponentially distributed packet inter-arrival time (represented by M), an arbitrary distribution for service time (represented by G) and one server.

Due to the contention medium, the overall problem can be seen as a set of interacting M/G/1 queues. We simplify the problem by developing an SMP model for the tagged vehicle that does not directly keep track of the queued requests but captures the channel contention and backoff behavior. This SMP model interacts with the M/G/1 queue hence we need to use fixed-point iteration to solve the overall model.

III. ANALYTIC MODELS

A. SMP model

The behavior of a tagged vehicle for packet transmission can be characterized using the SMP model in Figure 1.

The tagged vehicle is in *idle* state if there is no packet in its queue. After a packet is generated, the vehicle senses channel activity for DIFS time period. If channel is detected not busy during this period (with probability 1-q_b), the vehicle goes from *idle* state to *XMT* state, which means that a packet is transmitting. Otherwise, the vehicle will randomly choose a backoff counter in the range (0, W_0-1). The backoff counter will be decreased by one if channel is detected to be idle for a time slot σ (with probability 1-p_b), which is captured by the transition from state W_0-i to state W_0-i-1. If the channel is busy during a backoff time slot σ (i.e., another vehicle is transmitting a packet), the backoff counter of the tagged vehicle will be suspended and deferred for the duration of packet transmission time T, which presents the transition from state W_0-i to D_{W_0-i-1} with probability p_b. When the backoff counter reaches zero, the packet will directly be transmitted (an SMP transition occurs from state *0* to state *XMT* with probability one). In *XMT* state, a packet is transmitting. After the packet transmission, if there is no packet left in the queue of the tagged vehicle (with probability 1-ρ), the vehicle will go from *XMT* to *idle* state and wait for a new incoming packet. If there are packets left in the queue after a packet transmission (with probability ρ),

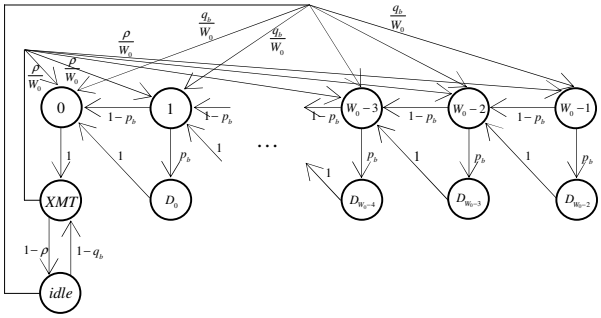


Figure 1. SMP model for IEEE 802.11 broadcast

the vehicle will sense the channel again for DIFS time and then randomly choose a backoff counter before transmitting the next packet.

Define the sojourn time in state j as T_j. The mean and variance of T_j in the SMP model are:

$$E[T_j] = \tau_j = \begin{cases} \sigma & j = 0,1,2,\cdots,W_0-1 \\ T & j = D_0, D_1, \cdots, D_{W_0-2} \\ T & j = XMT \\ \frac{1}{\lambda} + DIFS & j = idle \end{cases} \tag{2}$$

$$Var[T_j] = \theta_j^2 = \begin{cases} 0 & j = 0,1,2,\cdots,W_0-1 \\ Var[PA] & j = D_0, D_1, \cdots, D_{W_0-2} \\ Var[PA] & j = XMT \\ 1/\lambda^2 & j = idle \end{cases} \tag{3}$$

where T=$E[PA]$+T_H+$DIFS$+δ. The mean and variance of the packet length are $E[PA]$ and $Var[PA]$ respectively. T_H presents the packet header including physical layer header and MAC layer header. δ is the propagation delay. The parameters above are transferred to time unit for sojourn time calculation.

For the model in Figure 1, the embedded DTMC is first solved for its steady-state probabilities:

$$\begin{cases} v_j = (W_0 - j) \cdot v_{W_0-1} & j = 0,1,\cdots W_0-1 \\ v_{D_j} = (W_0 - j - 1) \cdot p_b \cdot v_{W_0-1} & j = 0,1,\cdots W_0-2 \\ v_{XMT} = \dfrac{W_0}{\rho + q_b(1-\rho)} \cdot v_{W_0-1} & \\ v_{idle} = \dfrac{(1-\rho)W_0}{\rho + q_b(1-\rho)} \cdot v_{W_0-1} & \end{cases} \tag{4}$$

$$v_{W_0-1}=\frac{2\left[\rho+q_b(1-\rho)\right]}{\left[W_0+1+p_b(W_0-1)\right]\left[\rho+q_b(1-\rho)\right]\cdot W_0+2(2-\rho)W_0} \quad (5)$$

Taking account of the mean sojourn time in each state, the steady-state probabilities of the SMP are given by:

$$\pi_i=\frac{v_i\tau_i}{\sum_j v_j\tau_j} \quad (6)$$

Therefore, the steady-state probability that a vehicle is in the *XMT* state is given by:

$$\pi_{XMT}=\frac{2T}{\left[\rho+q_b(1-\rho)\right]\left[(\sigma+p_b\cdot T)W_0+(\sigma-p_b\cdot T)\right]+2T+2(1-\rho)(1/\lambda+DIFS)} \quad (7)$$

Although the sojourn time in *XMT* state is T, the real packet transmission only occupies a portion of this sojourn time, which is $E[PA]+T_H+\delta=T-DIFS$. Hence, the probability that a vehicle transmits in steady state is $\pi_{XMT}(T\text{-}DIFS)/T$.

In Equation (7), three unknown parameters are:

- ρ: the probability that there are packets in the queue of the tagged vehicle.
- p_b: the probability that the channel is detected busy in one time slot by the tagged vehicle.
- q_b: the probability that the channel is detected busy in DIFS time by the tagged vehicle.

In Section III.C, we will see that p_b and q_b are functions of ρ, whereas ρ depends on the mean service time to transmit a packet. Therefore, the service time is derived first in the next subsection. Section III.C subsequently illustrates the relationships between these parameters and fixed point iteration algorithm is utilized to compute the numerical results for these parameters as well as the service time.

B. Service time computation

As mentioned above, each vehicle in the network can be modeled as an M/G/1 queue. The MAC layer service time is defined as the time interval from the time instant when a packet becomes the head of the queue and starts to contend for transmission, to the time instant when the packet is received.

The SMP model in Section III.A describes the behavior of a tagged vehicle continuously transmitting packets in its queue. In this section, the service time for any one packet in the queue needs to be derived. Therefore, the SMP model in Section III.A can be modified to contain an absorbing state as shown in Figure 2. By properly allocating the initial probability, the time to reach the absorbing state will be the service time for a packet transmission.

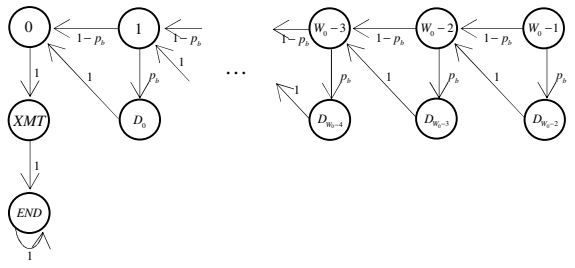


Figure 2. SMP model for the service time computation

For a newly generated packet in the tagged vehicle, if there is no packet in the queue and the channel is sensed idle for DIFS time with probability $(1-\rho)(1-q_b)$, the packet will be directly transmitted. In other words, the initial probability that a packet starts its transmission in *XMT* state is $(1-\rho)(1-q_b)$. Otherwise, the vehicle will randomly choose a backoff counter before the packet transmission. Therefore, the initial probability that a packet starts its transmission in state i ($i=0,1,\ldots,W_0-1$) is $[1-(1-\rho)(1-q_b)]/W_0$. Hence, the initial probabilities for all states are:

$$q_i=\begin{cases}\left[1-(1-\rho)(1-q_b)\right]/W_0 & for\quad i=0,1,...,W_0-1\\ (1-\rho)(1-q_b) & for\quad i=XMT\\ 0 & for\quad i=D_0,D_1,...,D_{W_0-2}\end{cases} \quad (8)$$

Since the Markov chain contains an absorbing state, the transition probability matrix can be partitioned so that:

$$P=\begin{bmatrix}Q & C\\ 0 & 1\end{bmatrix} \quad (9)$$

where Q is a $2W_0$ by $2W_0$ sub-stochastic matrix describing the probabilities of transitions only among the transient states. The fundamental matrix is:

$$M=(I-Q)^{-1} \quad (10)$$

Let X_{ij} be the random variable denoting the visit counts to state j before entering the absorbing state, given that embedded DTMC started in state i. The expected number of visits to state j starting from state i before absorption is given by the $(i,j)^{th}$ element of the fundamental matrix M. Hence,

$$E\left[X_{ij}\right]=m_{ij} \quad (11)$$

Due to the acyclic nature of the DTMC model in Figure 2, the fundamental matrix can be easily obtained through the definition of X_{ij} instead of computing (10).

$$M=[m_{ij}]=\begin{array}{c}\\ 0\\ 1\\ 2\\ \vdots\\ W_0-2\\ W_0-1\\ D_0\\ D_1\\ D_2\\ \vdots\\ D_{W_0-2}\\ XMT\end{array}\begin{array}{c}\begin{array}{cccccccccccc}0 & 1 & 2 & \cdots & W_0-2 & W_0-1 & D_0 & D_1 & D_2 & \cdots & D_{W_0-2} & XMT\end{array}\\ \begin{bmatrix}1&0&0&\cdots&0&0&0&0&0&\cdots&0&1\\ 1&1&0&\cdots&0&0&p_b&0&0&\cdots&0&1\\ 1&1&1&\cdots&0&0&p_b&p_b&0&\cdots&0&1\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ 1&1&1&\cdots&1&0&p_b&p_b&p_b&\cdots&0&1\\ 1&1&1&\cdots&1&1&p_b&p_b&p_b&\cdots&p_b&1\\ 1&0&0&\cdots&0&0&1&0&0&\cdots&0&1\\ 1&1&0&\cdots&0&0&p_b&1&0&\cdots&0&1\\ 1&1&1&\cdots&0&0&p_b&p_b&1&\cdots&0&1\\ \vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\ 1&1&1&\cdots&1&0&p_b&p_b&p_b&\cdots&1&1\\ 0&0&0&\cdots&0&0&0&0&0&\cdots&0&1\end{bmatrix}\end{array} \quad (12)$$

Furthermore, the variance of the number of visits can also be derived using the fundamental matrix. Define $M_D=[md_{ij}]$ by:

$$md_{ij}=\begin{cases}m_{ij} & if\quad i=j\\ 0 & otherwise\end{cases} \quad (13)$$

Define $M_2=[m_{ij}^2]$. Hence, the variance of the visit counts is:

$$\sigma^2=M\left(2M_D-I\right)-M_2 \quad (14)$$

The service time for a packet transmission starting from state i is given by:

$$S_i=\sum_j T_j\cdot X_{ij} \quad (15)$$

$$E[S_i]=E\left[\sum_j T_j\cdot X_{ij}\right]=\sum_j E\left[T_j\cdot X_{ij}\right]=\sum_j E\left[T_j\right]\cdot E\left[X_{ij}\right]=\sum_j \tau_j\cdot m_{ij}$$
$$=\begin{cases}(i+1)\sigma+i\cdot p_b\cdot T+T & for\quad i=0,1,\cdots,W_0-1\\ T & for\quad i=XMT\end{cases} \tag{16}$$

Since the sojourn time in state *0* is zero in the protocol instead of σ as specified in the model, we adjust the mean of S_i starting from i=0,1,…, W_0-1 by decreasing σ in the results. Hence, we obtain:

$$E[S_i]=\begin{cases}i\cdot\sigma+i\cdot p_b\cdot T+T & for\quad i=0,1,\cdots,W_0-1\\ T & for\quad i=XMT\end{cases} \tag{17}$$

The variance of S_i is given by (18).

$$Var[S_i]=Var\left[\sum_j T_j\cdot X_{ij}\right]=\sum_j Var\left[T_j\cdot X_{ij}\right]$$
$$=\sum_j\left\{Var\left[T_j\right]\cdot E\left[X_{ij}\right]+\left(E\left[T_j\right]\right)^2\cdot Var\left[X_{ij}\right]\right\}$$
$$=\sum_j\left(\theta_j^2\cdot m_{ij}+\tau_j^2\cdot\sigma_{ij}^2\right)$$
$$=\begin{cases}i\left\{Var[PA]\cdot p_b+T^2\cdot p_b\cdot(1-p_b)\right\}+Var[PA] & for\quad i=0,1,\cdots,W_0-1\\ Var[PA] & for\quad i=XMT\end{cases} \tag{18}$$

The service time for a packet transmission without conditioning on the starting state is presented as follows.

$$S=\begin{cases}S_0 & with\quad probability\quad \left[1-(1-\rho)(1-q_b)\right]/W_0\\ S_1 & with\quad probability\quad \left[1-(1-\rho)(1-q_b)\right]/W_0\\ \quad\vdots & \\ S_{W_0-1} & with\quad probability\quad \left[1-(1-\rho)(1-q_b)\right]/W_0\\ S_{XMT} & with\quad probability\quad (1-\rho)(1-q_b)\end{cases} \tag{19}$$

Therefore, the mean and variance of the service time are given by (20) and (21), respectively.

$$E[S]=\sum_i E[S_i]\cdot q_i=\frac{(\sigma+p_b\cdot T)\left[1-(1-\rho)(1-q_b)\right](W_0-1)}{2}+T \tag{20}$$

$$Var[S]=E\left[S^2\right]-\left(E[S]\right)^2=\sum_i E\left[S_i^2\right]\cdot q_i-\left(E[S]\right)^2$$
$$=\sum_i\left\{Var[S_i]+\left(E[S_i]\right)^2\right\}\cdot q_i-\left(E[S]\right)^2 \tag{21}$$

C. Fixed point iteration

In the previous section, the mean service time is shown to depend on three unknown parameters ρ, p_b and q_b, whereas ρ depends on the mean service time according to the M/G/1 queue behavior. Therefore, relationships between ρ, p_b and q_b are determined first in this section, and then the fixed point iteration algorithm is used to obtain the final solution.

Let N_{cs} denote the average number of vehicles in carrier sensing range of the tagged vehicle, whereas N_{tr} denote the average number of vehicles in transmission range of the tagged vehicle. Hence, without loss of generality, we have

$$N_{cs}=N_{tr}=2\beta R \tag{22}$$

The average number of vehicles in potential hidden area is:

$$N_{ph}=4\beta R-N_{cs}=2\beta R \tag{23}$$

From the tagged vehicle's point of view, p_b is the probability that it senses channel busy during one time slot in the backoff process. Since the channel is detected busy if there is at least one neighbor (i.e., a vehicle in the transmission range of the tagged vehicle) transmitting in a backoff time slot of the tagged vehicle, we have

$$p_b=1-\sum_{i=0}^{\infty}(1-P_{XMT})^i\frac{(N_{tr})^i}{i!}e^{-N_{tr}}=1-e^{-N_{tr}\cdot P_{XMT}} \tag{24}$$

where P_{XMT} is the probability that a neighbor is transmitting in a backoff time slot of the tagged vehicle, to be derived next.

Equation (7) shows that the probability that a vehicle transmits a packet in steady state is π_{XMT}(*T-DIFS*)/*T*. In addition, the time to transmit a packet is *T-DIFS*. Therefore, we can abstractly define the total time to be T_{total} as shown in Figure 3. Hence, π_{XMT}(*T-DIFS*)/*T* =(*T-DIFS*)/ T_{total}.

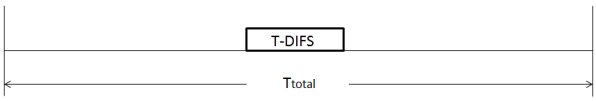


Figure 3. Abstraction of the packet transmission time

Suppose a neighbor of the tagged vehicle transmits a packet as shown in Figure 3 in time duration T_{total}, a backoff time slot of the tagged vehicle can occupy any one time slot within T_{total}.

For the first backoff time slot of the tagged vehicle, the time duration that can capture the transmission of the neighbor is *T-DIFS*+2σ. One extra time slot σ is the one just before transmission and another is the one just after transmission, which can capture the starting time instant and ending time instant of the packet transmission. Therefore, the probability that a neighbor's transmission is detected in the first backoff time slot of the tagged vehicle is π_{XMT}(*T-DIFS*+2σ)/*T*.

For a backoff time slot that is not the first backoff time slot of the tagged vehicle, the time duration that captures the transmission of the neighbor is 2σ, which captures the starting time instant of the transmission. This is because when the neighbor's transmission is detected in the first backoff time slot by the tagged vehicle, the backoff counter will suspend and wait until the end of this transmission for further decrement. Therefore, if the first backoff time slot detects the transmission, there is no chance for the later backoff time slots to detect the same transmission. As a result, the non-first backoff time slot can only detect the transmission when the starting point of the transmission falls within this time slot. Therefore, the probability that a neighbor's transmission is detected in non-first backoff time slot of the tagged vehicle is $\pi_{XMT}\cdot 2\sigma/T$.

Since the probability that a backoff time slot is the first backoff time slot is $1/W_0$ and non-first backoff time slot is $(1-1/W_0)$, the probability that a neighbor's transmission is detected by a backoff time slot of the tagged vehicle is:

$$P_{XMT}=\frac{1}{W_0}\cdot\frac{T-DIFS+2\sigma}{T}\pi_{XMT}+\left(1-\frac{1}{W_0}\right)\cdot\frac{2\sigma}{T}\pi_{XMT} \tag{25}$$

Next, q_b denotes the probability that channel is detected busy by the tagged vehicle in the DIFS duration. Therefore, we can similarly define P_{XMT}' to be the probability that a neighbor's transmission is detected in the DIFS duration by the tagged vehicle.

 ISBN: 978-1-61208-126-7

$$P_{XMT}' = \frac{T - DIFS + 2DIFS}{T}\pi_{XMT} = \frac{T + DIFS}{T}\pi_{XMT} \quad (26)$$

Hence, q_b is given by (27).

$$q_b = 1 - \sum_{i=0}^{\infty}(1 - P_{XMT}')^i \frac{(N_{tr})^i}{i!} e^{-N_{tr}} = 1 - e^{-N_{tr} \cdot P_{XMT}'} \quad (27)$$

Based on Equations (24)-(27), q_b is expressed in terms of p_b to simplify the iteration algorithm:

$$q_b = 1 - (1 - p_b)^{\frac{(T+DIFS)W_0}{T-DIFS+2\sigma W_0}} \quad (28)$$

From the above analysis of the relationship between two parameters ρ and p_b (q_b can be expressed in terms of p_b), we notice that p_b depends on ρ and p_b itself. Hence, we denote $p_b=g(\rho,p_b)$ and the reciprocal of mean service time for M/G/1 queue to be $\mu=h(\rho,p_b)$. The fixed point iteration algorithm is outlined as follows according to the import graph in Figure 4.

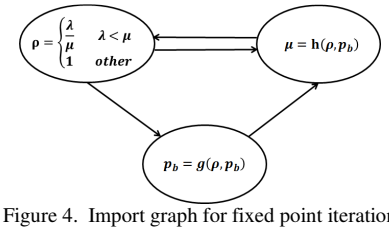


Figure 4. Import graph for fixed point iteration

Fixed point iteration algorithm:

Step 1: Initialize ρ=1, which is the saturation condition;

Step 2: With ρ, solve p_b according to (24)(25)(7)(28);

Step 3: With ρ and p_b, calculate service rate $\mu=1/E[S]$ according to (20);

Step 4: If $\lambda<\mu$, $\rho=\lambda/\mu$; otherwise, ρ=1;

Step 5: If ρ converges to the previous value, then stop the iteration algorithm; otherwise, go to step 2 with the updated ρ.

By utilizing the fixed point iteration algorithm, the parameters ρ, p_b, q_b, π_{XMT} as well as the mean and the variance of the service time can be determined, which are used for performance indices computation in the next section.

IV. PERFORMANCE INDICES

A. Transmission delay

The packet transmission delay is defined as the average delay a packet experiences from the time the packet is generated, until the time the packet is successfully received by all neighbors of the vehicle that generates the packet. The transmission delay $E[D]$ includes the queuing delay and medium service time (due to backoff, packet transmission, and propagation delay, etc.)

The expected queuing delay can be obtained from the *Pollaczek-Khinchin* mean value formula of M/G/1 queue:

$$E[D_q] = \frac{\lambda\left(Var[S] + (E[S])^2\right)}{2(1 - \lambda E[S])} \quad (29)$$

The average packet transmission delay is then calculated as:

$$E[D] = E[D_q] + E[S] \quad (30)$$

B. Packet Delivery Ratio

The PDR is defined as: given a broadcast packet sent by the tagged vehicle, the probability that all vehicles in its transmission range receive the packet successfully. Taking account of hidden terminals, we have

$$PDR = P(N_{cs})P(N_{ph}) \quad (31)$$

where $P(N_{cs})$ is the probability that no vehicles in the transmission range of the tagged vehicle transmits when the tagged vehicle starts transmission, and $P(N_{ph})$ is the probability that no transmissions from the vehicles in the potential hidden terminal area collide with the broadcast packet from the tagged vehicle.

$P(N_{cs})$ can also be interpreted as the no-concurrent transmission probability, i.e., two packets do not start transmission at the same time. Since DCF employs a discrete-time backoff scheme, if the backoff process is involved, a vehicle is only allowed to transmit at the beginning of each slot time after an idle DIFS. Therefore, if the tagged vehicle has not gone through the backoff process before transmitting the packet (with probability $(1-\rho)(1-q_b)$), the concurrent transmission will not occur. Otherwise, the packet transmission is synchronized to the beginning of a slot time, and concurrent transmission may occur if other vehicles' transmission is also synchronized by the backoff process. From the model, we know that the probability that a neighbor starts to transmit a packet at the beginning of the same time slot with the tagged vehicle is $\pi_0 = \pi_{XMT} \cdot \sigma/T$. This is because the sojourn time in state *0* is one time slot σ as shown in the SMP model, hence, π_0 is the probability that a vehicle starts to transmit in the beginning of a time slot immediately after the backoff process. Hence, $P(N_{cs})$ is:

$$\begin{aligned} P(N_{cs}) &= \left[1 - (1-\rho)(1-q_b)\right] \cdot \left[\sum_{i=0}^{\infty}(1-\pi_0)^i \frac{(N_{cs}-1)^i}{i!} e^{-(N_{cs}-1)}\right] \\ &\quad + (1-\rho)(1-q_b) \\ &= \left[1 - (1-\rho)(1-q_b)\right] \cdot e^{-(N_{cs}-1)\pi_0} + (1-\rho)(1-q_b) \end{aligned} \quad (32)$$

Since the transmission time for a packet is $T\text{-}DIFS=E[PA]+T_H+\delta$, the event that transmission from hidden terminals collides with the tagged vehicle's transmission only happens when hidden terminals start to transmit during the vulnerable period $2(T\text{-}DIFS)=2(E[PA]+T_H+\delta)$ [6]. Using $\pi_{XMT}=T/T_{total}$ as an abstraction of the steady state behavior shown in Figure 3, the probability that a vehicle starts to transmit during the vulnerable period is:

$$\frac{2(T - DIFS)}{Ttotal} = \pi_{XMT}\frac{2(T - DIFS)}{T} \quad (33)$$

Therefore,

$$\begin{aligned} P(N_{ph}) &= \sum_{i=0}^{\infty}\left(1 - \pi_{XMT} \cdot \frac{2(T - DIFS)}{T}\right)^i \frac{(N_{ph})^i}{i!} e^{-N_{ph}} \\ &= e^{\frac{-2(T-DIFS)\cdot N_{ph}\cdot\pi_{XMT}}{T}} \end{aligned} \quad (34)$$

V. NUMERICAL AND SIMULATION RESULTS

The computation for analytic models and corresponding simulations are conducted in Matlab. Table 1 shows the parameters used in this paper, which reflect typical DSRC

network settings in [1]. Figures 5 and 6 present the mean transmission delay and PDR, respectively, vs. the vehicle density β (# vehicles per meter), data rate R_d (Mbps), packet arrival rate λ (packets per second) and average packet length $E[PA]$ (bytes).

Table 1. Parameters

Parameters	Values	Parameters	Values
Tx range R	500 m	Propagation delay δ	0 μs
Average Packet Length $E[PA]$	variable	Variance of Packet Length $Var[PA]$	0
PHY preamble	40 μs	PLCP header	4 μs
MAC header	272 $bits$	CWMin W_0-1	15
Packet arrival rate λ	variable	Vehicle density β	variable
Slot time σ	16 μs	DIFS	64 μs

Figures 5 and 6 show that the analytic results from the model have better match with the simulation results than those from the model in [6]. The 95% confidence intervals are shown in the figures.

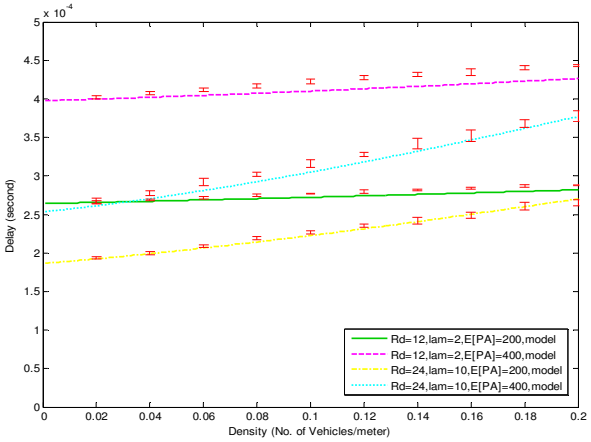


Figure 5. Delay of DSRC Highway safety messaging

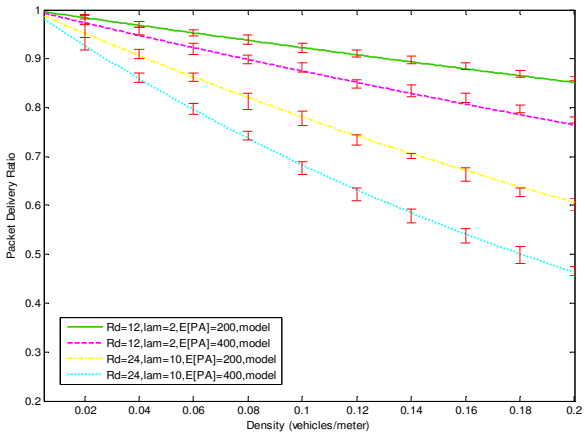


Figure 6. PDR of DSRC Highway safety messaging

Since the SMP model considers the fact that a packet can be directly transmitted without undergoing backoff process, the delay is lower compared with [6]. Another observation in Figure 5 is that high data rate and shorter packet length facilitate the decrease of the transmission delay.

The PDR decreases fast as the density β increases as shown in Figure 6. Similar to the delay, PDR also benefits from high data rate and short packet length.

VI. CONCLUSION AND FUTURE WORK

In this paper, a more general and accurate analytic model using SMP has been developed to characterize the behavior of DSRC for highway safety communications. The model is cross validated against simulations. Moreover, the performance with different input parameters is analyzed to suggest better parameter settings that will improve the performance by decreasing the delay and increasing PDR. In future, performance optimization will be conducted for more parameters including W_0. The tradeoff between delay and PDR will be evaluated based on optimization results. In addition, the SMP model will be extended to incorporate different packet arrival processes such as Markov modulated Poisson process (MMPP), Markov arrival process (MAP) instead of Poisson arrival. Besides one-hop direct broadcast transmission strategy, multi-hop and multi-cycle transmission strategy will also be considered in future.

ACKNOWLEDGMENT

The authors would like to thank Yang Zhao for his help with simulation and NSF grants (CNS-1018605 and CNS-1017722) to support this research.

REFERENCES

[1] ASTM E2213-03, "Standard Specification for Telecommunications and Information Exchange Between Roadside and Vehicle Systems 5GHz Band Dedicated Short Range Communications (DSRC) Medium Access Control (MAC) and Physical Layer (PHY) Specifications", ASTM Intl. Jul, 2003.

[2] J. Yin, T. Elbatt, et al., "Performance evaluation of safety applications over DSRC vehicular ad hoc networks", *ACM VANET*, pp. 1-9, 2004.

[3] M. T. Moreno, D. Jiang, and H. Hartenstein, "Broadcast reception rates and effects of priority access in 802.11-based vehicular ad-hoc networks", *ACM VANET*, pp. 10-18, 2004.

[4] Q. Xu, T. Mak, J. Ko, and R. Sengupta, "Vehicle-to-vehicle safety messaging in DSRC", *ACM VANET*, pp. 19-28, 2004

[5] X. Ma, X. Chen, and H. H. Refai, "Unsaturated performance of IEEE 802.11 broadcast service in vehicle-to-vehicle networks", *IEEE VTC*, pp. 1957-1961, 2007.

[6] X. Chen, H. H. Refai, and X. Ma, "A Quantitative Approach to evaluation DSRC Highway Inter-vehicle Safety Communication", *IEEE GLOBECOM*, pp. 151-155, 2007.

[7] X. Ma, X. Chen, and H. H. Refai, "Performance and reliability of DSRC vehicular safety communication: a formal analysis", *EURASIP J. on Wireless Comm. and Networking*, pp. 1-13, 2009.

[8] D. Wang, R. Fricks, and K. S. Trivedi, "Dealing with non-exponential distributions in dependability models", *Performance Evaluation - Stories and Perspectives*, G. Kotsis, editor, pp. 273-302, 2003.

[9] K. S. Trivedi, *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, Second Edition, John Wiley, 2002.

[10] L. Tomek and K.S. Trivedi, "Fixed-Point Iteration in Availability Modeling", *Informatik-Fachberichte*, Vol. 283, M. DalCin (ed.), pp. 229-240, Springer-Verlag, Berlin, 1991.

[11] IEEE 802.11 Working group, "Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications", ANSI/IEEE Std. 802.11, Sep. 1999.

 ISBN: 978-1-61208-126-7

A Resource Management Strategy to Support VoIP across Ad hoc IEEE 802.11 Networks

Janusz Romanik
Radiocommunications Department
Military Communications Institute
Zegrze, Poland
j.romanik@wil.waw.pl

Piotr Gajewski, Jacek Jarmakiewicz
Faculty of Electronics
Military University of Technology
Warsaw, Poland
{pgajewski, jjarmakiewicz}@wel.wat.edu.pl

***Abstract* — This paper describes the concept of the resource management strategy in ad hoc networks for rescue operations. The presented strategy is a result of the new outlook on the IEEE 802.11 networks capabilities and performance enhancement. The proposed solution is dedicated to real time services support and is based on the concept of the Resource Manager that organizes and controls the whole traffic in the network. Novel procedures were developed and applied in order to organize the network and manage the real time traffic. A method of the available bandwidth measurement and estimation was introduced. Large scale simulations for different numbers of Voice over IP (VoIP) sources and various voice codecs have been carried out. They show the increase of channel utilization reaching over 80% and significant growth of the network capacity.**

Keywords - IEEE802.11 WLANs, ad-hoc networks, VoWiFi, resource management

I. INTRODUCTION

For over ten years a permanent development of IEEE 802.11 Wireless Local Area Networks (WLANs) is being observed [1]. Among the many advantages they offer, users appreciated the convenience and simplicity when accessing the network and establishing high data rate wireless connection. Thanks to a low cost and a small size of devices, nowadays they seem too ubiquitous. WLAN drivers are embedded in many different devices like notebooks, mobile phones, Personal Data Assistants (PDAs), cameras, etc. Despite the fact that WLANs were originally designed for data transport, today it is also demanded of them to be efficient for real time services support.

Another advantage of WLANs results from the ad hoc mode, which is a method for wireless devices to directly communicate with each other. Operating in ad hoc mode allows all wireless devices within each other's range to discover and communicate in a peer-to-peer manner without involving the central access point. This mode offers mobility and communications between users in areas without infrastructure or in all places with damaged infrastructure. From this point of view, WLANs operating in ad hoc mode can be a very promising solution for users, such as the fire brigade, rescue team, police squad or small military unit [2,3]. The possible scenario is to use the ad hoc network for public-safety or search-and-rescue operations.

An important issue for such network is the ability to support cooperation between two or more emergency services, e.g., the fire brigade, police squad, rescue team, medical service. On the other hand, it must be stressed that the performance of the network decreases as the number of wireless users grows. For that reason, a smart mechanism should be introduced, which allows topology control and network scalability [4].

The effect of the hidden node is one of the most difficult problems to solve, because it is intrinsic to the nature of the WLANs. The RTS/CTS mechanism is not recommended for the transmission of small packets, e.g. VoIP. A possible solution is to use an additional signaling channel, however it requires changes in the physical layer. This issue was widely discussed in [14,15].

When considering the hierarchical structure of the command system of the emergency services, different ranks of users should be taken into account. This will affect the priority of users, as well as the type of allowed services.

Among many wireless solutions, IEEE 802.11 networks seem to be the most popular. Although the most common weakness of WLANs is the insufficient support of the real time services [5,6], the authors formulated a new outlook on the IEEE 802.11b network capability and possible performance enhancement. Despite the fact that there is a wide range of WLANs specifications, the issue of network optimization still remains open. QoS mechanisms were the subject of the IEEE 802.11e standard [7]. However, these mechanisms cannot guarantee the quality of services, although they slightly improve the network efficiency [8].

The voice capacity of IEEE 802.11 networks is gaining increasing attention in the literature. Methods of VoWiFi optimization, including voice codec negotiation, audio packets aggregation as well as the MAC protocol adaptation, can be found in many papers. In [8], the influence of the MAC protocol on the network performance was shown. This protocol operates in contention mode and thus inevitably introduces the PHY layer overheads, Backoff and protective periods, ACK frames and retransmissions in some cases. In [9], authors analyzed the effect of the coding rate and packet

 ISBN: 978-1-61208-126-7

size on the voice capacity of the Distributed Coordination Function (DCF).

In [10], dynamic CW adaptation was suggested in order to minimize the number of collisions. The idea of the voice coding bit rate adaptation to the available network bandwidth was described in [22]. Results of experiments confirmed the efficiency of the new scheme. The impact of different configuration parameters on the ad-hoc network performance was presented in [23]. Following parameters were analyzed, the type of codec, packetization interval and the data rate. In [24], authors presented the results of the capacity measurement of the IEEE 802.11e network for each access category. They also analyzed the effect of the TCP traffic on VoIP streams. In conclusion, they stated that 802.11e standard can protect the quality of VoIP if there is TCP traffic added. However, it can not improve the capacity of the network.

Although proposed methods can improve network efficiency, the question as to how to guarantee the quality of services still remains open [13]. Furthermore, there is still a lack of an efficient Call Admission Control (AC) mechanism [12,13]. The present article is an attempt to fill this gap.

This paper presents the general concept of the resource management strategy and provides information on introduced procedures. All proposed mechanisms are connected with each other and interact within a individual device as well as within the whole network.

The rest of the paper deals with the concept and assumptions (Section 2), the description of the proposed mechanisms (Section 3), simulation results and their discussion (Section 4), conclusions (Section 5) and future work (Section 6).

II. CONCEPT AND ASSUMPTIONS

In the case under consideration, the aim of the network optimization is to get as high as possible number of VoIP streams with guaranteed voice quality. The assumed network operates in ad hoc mode and consists of small group of users, e.g., fire brigade or rescue team. In emergency situations they typically use voice communication. Therefore the authors made an assumption that there is only one type of service, namely VoIP.

Users have different ranks, which determines some differences between priorities. Thus, the trade off between the available bandwidth, the allowed number and the rank of users is introduced intentionally.

The network model assumes WLAN based solution. The authors decided to use the IEEE 802.11b standard as offering good throughput and modulations more resistant to interferences, which is a real advantage of the network operating in ad hoc mode. MAC QoS mechanisms defined in the IEEE 802.11e standard were also taken into account. These mechanisms are a good starting point to enable prioritization and bandwidth reservation in ad hoc network [12,13]. In particular, the authors introduced the adaptation of a Contention Window (CW) size to the type of frame and the rank of the user.

Another issue concerns the optimal balance between the traffic load and the services quality in ad-hoc networks.

It is expected that the proposed range of adaptation and introducing of new mechanisms will not demand a high cost of implementation and will be feasible.

Fig. 1 illustrates the concept of efficiency improvement of WLAN for VoIP support. The available bandwidth level is the main factor allowing assessment of the traffic load in the network. Cross-layer mechanisms are crucial for network performance improvement. They enable the RT traffic shaping or MAC adaptation if the available bandwidth is too small or if the level of service is not satisfactory. The CAC mechanism prevents new VoIP calls if the available bandwidth level is too low.

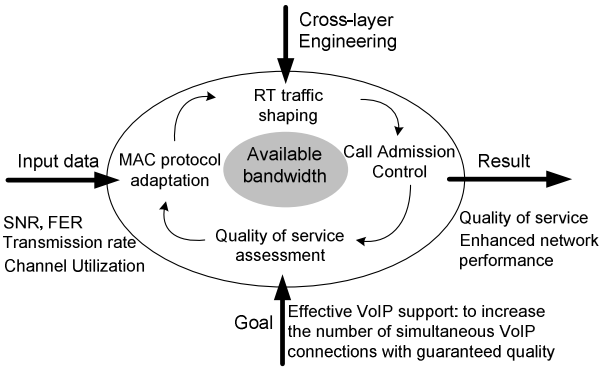


Figure 1. Performance enhancement of WLAN for VoIP support.

In the proposed solution, the network consists of different rank users. For the sake of simplicity, high rank users shall be denoted as special users while the rest shall be referred to as commercial users. It perfectly corresponds to the scenario of the humanitarian aid, when volunteers help people in service. Another example can be the situation when the fire brigade, police and civilians cooperate within small groups while strengthening an embankment during a flood. However, if the available bandwidth is too small, special users prevail over the network. Eventually, the lowest rank users can be completely blocked.

A separate question is how to assess the resources of the network, e.g. channel utilization. Since ad-hoc networks are bandwidth limited, not all measurement methods can be applied [15-17].

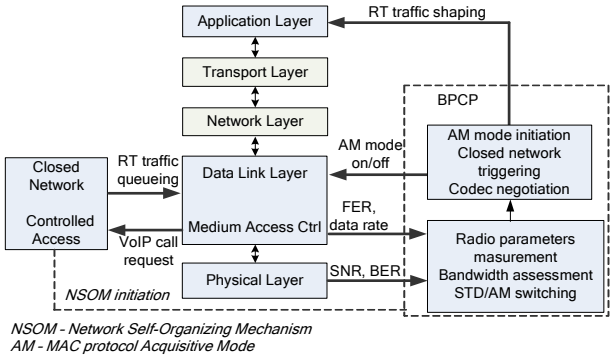


Figure 2. BPCP alignment with protocol stack.

Fig. 2 illustrates the extended protocol stack with cross-layer interactions. Bandwidth Prediction Control Protocol (BPCP) allows monitoring of parameters in the physical layer, to measure the channel utilization level and also to switch MAC protocol states, as explained in subsequent sections. RT traffic shaping relays on codec negotiation and audio packets aggregation. Closed Network Mode is based on the concept of the Resource Manager that controls traffic in the network.

A. Bandwidth Estimation

The available bandwidth is crucial for optimization of the Wi-Fi ad-hoc network. Therefore, the authors proposed to implement BPCP that enables to measure the channel utilization level and to estimate the available bandwidth.

BPCP takes advantage of WLAN card drivers that enable the measurement of SNR in the PHY layer and BER calculation, and passing these parameters to the Data Link Layer. If nodes operate in promiscuous mode, they can receive all the traffic sent across the network. As a result, the bandwidth utilization is assessed in all nodes of the network independently and continuously for predefined periods called Sampling Intervals.

From the PHY layer point of view, stations can detect the channel state (idle or busy - which means transmission) and if they operate in promiscuous mode, they can receive and process all frames. The type of received frames (RTS, CTS, DATA, ACK) is recognized in the data link layer.

Knowing the bit rate and the length of received frames it is possible to calculate their transmission duration in the radio channel, denoted as t_{AF} in (1).

$$t_{AF} = \frac{AF\_length\,(bits)}{AF\_bitRate\,(bits/\sec)} \quad (1)$$

Having knowledge of t_{AF} parameters, it is then possible to determine the channel utilization coefficient for the interval, e.g. from t_1 to t_2

$$U = \frac{t_{d1} + t_{d2} + 2 \cdot t_{ACK} + 2 \cdot SIFS + 2 \cdot DIFS}{t_2 - t_1} \quad (2)$$

where: t_{d1}, t_{d2} denotes the duration of the first and second data frames; t_{ACK} represents the ACK frame duration, Fig. 3.

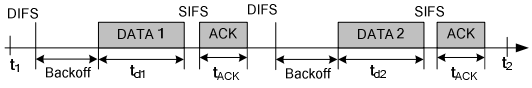


Figure 3. Transmission scheme in a contention mode of WLAN.

When the current and the previous channel utilization is estimated, BPCP makes forecasts for the next period.

Fig.4 shows the extended WLAN sublayer of the mobile node. This sublayer contains Throughput Meter In and Throughput Meter Out components to measure all incoming and outgoing traffic. This information is used to assess the total traffic load as well as the available bandwidth.

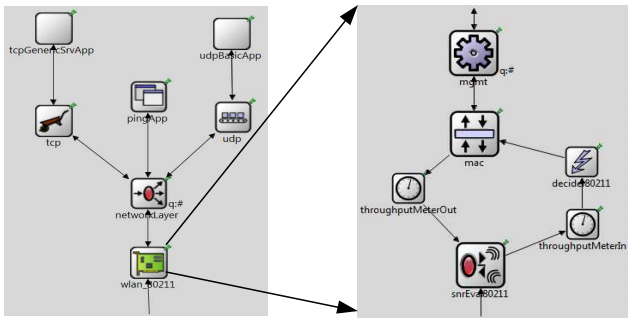


Figure 4. The extended WLAN sublayer of the mobile node.

To assess the network throughput in a contention mode, theoretical analysis was performed and simulations were made using the OMNET++ v4.0 simulation tool.

For the purposes of analysis and simulation, the following parameters were assumed: G.711 voice codec; typical protocol headers (MAC header = 30B, IPv4 header =20B, UDP header = 8B); free space propagation model and lack of mobility. The issue of mobility is crucial for NRM determination and is the topic of further study.

The values of the MAC parameters are listed in Table I.

TABLE I. MAC PARAMETERS

Parameter	Value
DIFS	50 µs
SIFS	10 µs
Slot Time	20 µs
CWmin	32
CWmax	1023
Data Rate	2Mbit/s
PHY header	192 µs
MAC header	34 bytes
ACK	304 µs

The main attributes of the G.711 codec are shown in Table II.

TABLE II. G.711 CODEC CHARACTERISTICS

Codec	G.711
Bit rate [kbit/s]	64
Framing interval [ms]	20
Payload [B]	160
Packets/sec	50

The results of the simulation are presented in Fig. 5. Normal distribution of a throughput estimator was assumed, as well as a confidence interval with α=0.1 and β=1.64 (for cumulative distribution function equal to 0.9).

The period of time required for the transmission of one data frame and the acknowledging frame takes nearly 1,8ms. For that reason it is possible to send 11 acknowledged frames during one second. Audio packets are generated by codec periodically every 20ms. Assuming that stations work synchronously, i.e., after the first one had transmitted a packet, the second one generates it, then it is possible to

 ISBN: 978-1-61208-126-7

obtain the network throughput equal to 1,3Mb/s, Fig. 5. Higher traffic load will cause an increase of the collision rate and a drop in network efficiency.

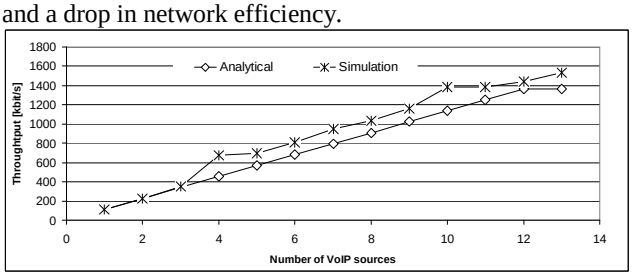


Figure 5. Wi-Fi network throughput - contention mode, data rate 2Mbit/s.

B. MAC Protocol States

At the beginning of the operation, special stations can cooperate with commercial and use standard access schemes, until BPCP detects the insufficient bandwidth and initiates the Acquisitive Mode (AM).

During AM mode, the Backoff interval is minimized according to the rank of the user. As a result, special stations prevail over the network. Only a small part of the bandwidth can be hard-won by remaining users. To determine the Backoff interval, the Contention Window parameter is used, however different values have been introduced, depending on the rank of the user and the type of frame (Control, Data, Broadcast or RTData).

If the available bandwidth is still too small, BPCP triggers a mechanism called the Network Self-Organizing Mechanism, which is responsible for creating a Closed Network Mode. From this moment on, Wi-Fi network operates in a point-coordinated mode. Fig. 6. presents the states of MAC protocol for the proposed protocol extension

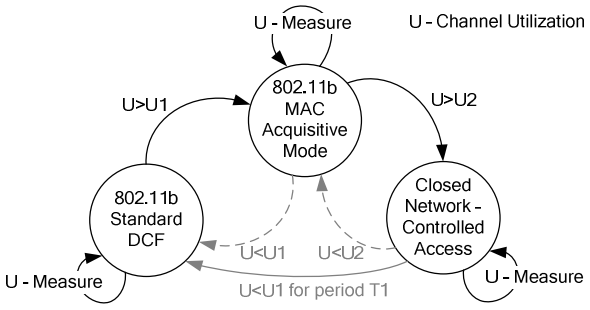


Figure 6. MAC protocol states.

An important issue is to determine the proper level of channel utilization for triggering between MAC AM and Closed Network Mode. To resolve this problem, the authors applied the Pareto optimization approach.

Simulation results obtained for 2Mbit/s data rate and G.711 voice codec are presented below.

Fig. 7 shows the network throughput vs. traffic load. Triggering levels are also presented. If the throughput reaches limit denoted by B, the station switches from standard mode to AM. If it reaches another limit denoted by C, the station switches to Closed Network Mode. Fig. 8 presents collisions vs. traffic load. The critical level of collisions is denoted by A. This information may be used additionally by BPCP.

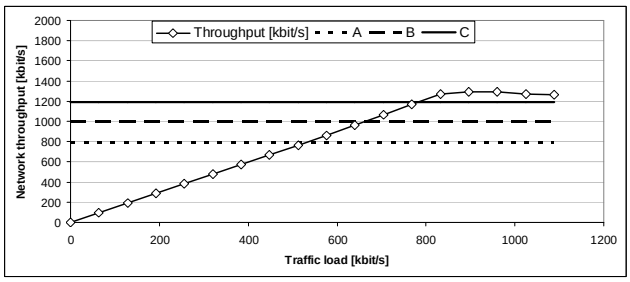


Figure 7. Network throughput vs. traffic load (where triggering levels are denoted as follows: A - switch from AM Mode to std., B - switch from std. to AM Mode, C - switch to Closed Network Mode).

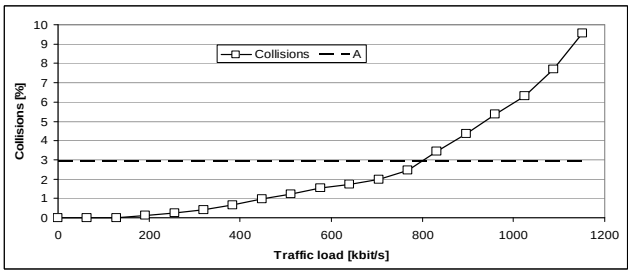


Figure 8. Collisions vs. traffic load (where A denotes the critical level of collisions).

Results of simulations performed in order to estimate the acceptable number of VoIP connections, depending on the type of voice codec and MAC protocol parameters in a contention mode, were widely discussed in literature [9,10,19,20]. However, the question where and how to implement the AC mechanism and how to manage the traffic in the network still remains open. The AC mechanism is necessary to prevent new calls if there is not enough bandwidth. In a contention mode, stations are not aware of the traffic load and try to transmit frames every time they have a packet to send. For this reason, a Closed Network Mode was proposed with a station named the Network Resource Manager (NRM) that manages the network.

III. CLOSED NETWORK MODE

A. Network Self-Organizing Mechanism

At the beginning, all stations work in a contention mode with standard parameters, Fig. 9. In the background, Neighbor Discovery Procedure is performed, which is based on broadcasting Neighbor Request and Neighbor Response frames [21]. This procedure allows recognition of the surroundings by collecting data from other nodes, namely: received signal strength and noise, battery level and rank of the station. Based on this information, each station determines its own *NRM Readiness* coefficient, which describes whether the station is ready to play a network manager role. This mechanism is still under implementation in OMNET++ v4.0.

 ISBN: 978-1-61208-126-7

If BPCP again detects the insufficient bandwidth coincidence, a station changes the mode to AM, while Neighbor Discovery Procedure is still in the background, Fig. 9. When a first station detects the insufficient bandwidth, it initiates Network Self-Organizing Mechanism (NSOM). Only stations with a certain *NRM Readiness* coefficient are allowed to participate in this phase. If necessary, information on the network topology is refreshed by sending Neighbour Request frame, which contains the last *NRM Readiness* coefficient of the sending station. When these frames are exchanged, the station with the highest coefficient sends a Request for RT frame. From this moment on, the network operates in a Closed Network Mode and all traffic is controlled by the resource manager till *NRM Timeout* elapses and the procedure for NRM Determination starts again, Fig. 10.

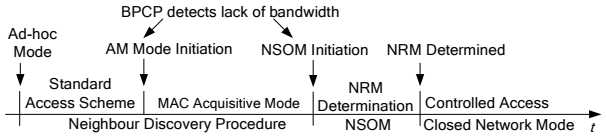


Figure 9. Network Self Organizing procedure.

B. *Real Time Traffic Management*

When the NRM station is determined, it sends a NRM Request broadcast frame informing that nodes are allowed to call for a bandwidth reservation. Some stations respond with RT Confirm frames if they have RT packets to send. The NRM Request frame is sent periodically to disseminate the list of queued stations and also the current queue limit, Fig. 10. A more detailed description of the algorithm can be found in [18].

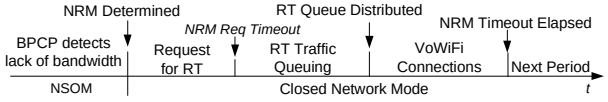


Figure 10. RT traffic management.

If the queue limit is reached or *NRM Req Timeout* has elapsed, the NRM station sends a RT Queue frame containing:

- queue size: number of STAs in queue,
- number of cycles: number of queue repetition,
- voice codec type,
- data rate,
- MAC address and order of stations in the queue.

After receiving the RT Queue frame, the first station on the list is allowed to transmit after DIFS and receives an ACK frame after SIFS, Fig. 11. The next station in queue transmits data frame after DIFS. The number of cycles describes how long nodes will transmit data in a given order.

After each transmission of DATA and ACK, stations decrease their *TransmissionIndex* and are allowed to send after it reaches zero. After a predefined number of cycles, the NRM station again sends a NRM Request frame to give a chance to transmit for stations that were out of queue during the preceding period.

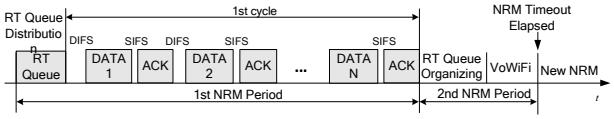


Figure 11. RT traffic queue.

An unpredictable NRM termination may occur, e.g. as a result of depletion of the battery, which should be taken into account. In such a situation nodes will detect a lack of frames from NRM for the assumed timeout. Since this moment on, the station with the second highest *NRM Readiness* coefficient starts playing this role.

In order to organize a closed network and manage RT traffic, the following management frames were introduced:

- Neighbor Request and Neighbor Response - for neighborhood discovering,
- NRM Request - for initiation of the RT traffic queuing phase,
- RT Confirm - for the bandwidth reservation,
- RT Queue - distribution of RT traffic queue.

The detailed description of the management frames structure can be found in [18]. Because all of these frames are of a broadcast type, all receiving stations are forced to process it in the data link layer, although acknowledgement is not sent. The structure of new frames is the same as defined in the IEEE 802.11 standard for management frames and consists of MAC Header and Frame Body containing information fields. The maximum size and capacity of frames are presented in Table III.

TABLE III. MANAGEMENT FRAMES SIZE AND CAPACITY

Frame Type	Frame max size [B]	Number of addresses
NRM Request	240	35
RT Confirm	40	1
RT Queue	280	35

IV. VOIP CAPACITY ANALYSIS

In order to assess the time required to organize the RT traffic, analytical investigations were performed. It was assumed that NRM is determined, avg. Backoff is equal to 100μs and 10 nodes compete for bandwidth reservation, Fig. 12.

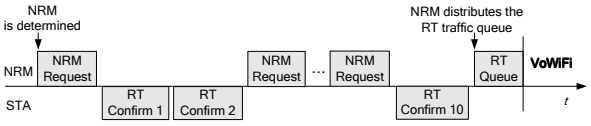


Figure 12. RT traffic scheduling procedure.

The size and the amount of frames exchanged in this procedure are presented in Table IV.

TABLE IV. AVERAGE SIZE AND NUMBER OF EXCHANGED FRAMES

Frame type	Frame avg. size [B]	Frames number
NRM Request	100	5
RT Confirm	40	10
RT Queue	100	1

If the data rate is set to 1Mbit/s, one cycle required to schedule the RT traffic takes approximately 14ms and this period reaches 10ms if the data rate increases to 2Mbit/s. Assuming some collisions, this duration should not exceed 20ms.

Synchronous RT data transmission in a Closed Network Mode can be verified by using an analytical as well as simulation model. For the sake of convenience, e.g. in order to apply different input parameters, the authors used COMNET 3 simulation tool.

The aim of simulations was to assess the channel utilization and the number of possible simultaneous VoIP calls as a function of the data rate. The following assumptions were made:

- network stations with commercial voice codec (G.711) with attributes defined in Table I,
- MAC/PHY parameters: SIFS = 10μs, DIFS = 50μs, PLCP Header + Preamble = 192μs,
- packets with standard protocol headers: MAC = 30B, IPv4 =20B, UDP = 8B.

The channel utilization vs. the number of VoIP calls and various data rates was shown in Fig. 13 and Fig. 14.

In the phase of synchronous RT data transmission, there are only two cases when the channel is idle: DIFS which precedes data frame transmission and SIFS between data and ACK frames.

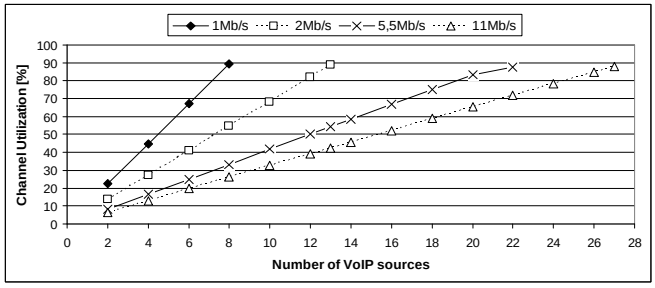


Figure 13. Channel utilization vs. number of VoIP streams for G.711 voice codec (64kbit/s).

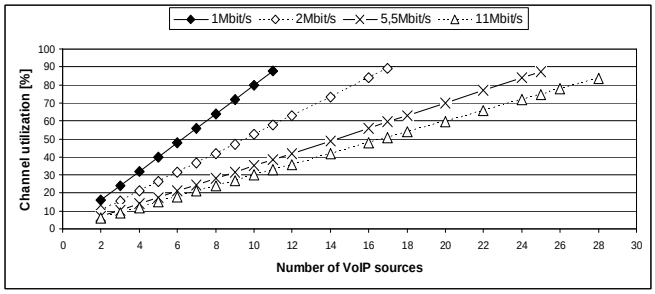


Figure 14. Channel utilization vs. number of VoIP streams for G.726 voice codec (32kbit/s).

An increasing number of VoIP connections leads to a linear growth of channel utilization, up to 90%. Better channel utilization is unachievable. This is a result of the fact that although the number of frames sent in a given period increases for higher data rates, there are still constant idle periods that separate frames.

The delay of RT packets results from the data rate and the sequence number of a given station in the whole queue. Thus, this delay does not exceed two dozens of milliseconds. When the data rate grows, the time needed for transmission of one frame becomes shorter, while DIFS and SIFS remain on the same level. Therefore it is possible to set up more VoIP connections, however the channel utilization cannot exceed 90%.When G.711 codec is used and the data rate is set up to 11Mb/s, up to 27 VoIP calls are available.

V. CONCLUSIONS

We have presented the concept of the resource management strategy in ad-hoc networks for rescue operations. This strategy is a result of the new outlook on the 802.11 WLANs capabilities and performance enhancement.

A set of novel procedures was developed with a view of organizing the network and managing the real time traffic. These procedures were validated analytically and by simulations, and results were included. The proposed method of the available bandwidth measurement and estimation works correctly.

The procedure of RT data synchronous transmission in a Closed Network Mode was verified by simulation. Results of tests allowed estimating the channel utilization achieving over 80% when synchronous transmission was applied. If the number of stations in a queue is set correctly, the delay of the RT data frame transmission is limited to two dozens of milliseconds and results mainly from the data rate.

The presented results were obtained under the assumption that only UDP traffic is transferred across the network. The impact of the TCP flows on the network performance requires further analysis.

The proposed mechanisms were developed as a result of a completely new approach to the support of RT data transmission in 802.11 ad-hoc network. They enrich standard procedures and enable an efficient utilization of the channel.

VI. FUTURE WORK

In this article, we have only presented the resource management strategy to support VoIP traffic. We described the procedures enabling the organizing of the network and real time traffic management.

For future research it would be interesting to study the effect of the TCP traffic on the network capacity for VoIP. Based on this work, we are going to investigate how to efficiently manage the network where VoIP streams are combined with the TCP flows.

The issue of nodes mobility is crucial for NRM determination and will be the topic of further study.

Furthermore, we would like to devote attention to the aspect of the distributed network management. This includes optimization of the scheme for determining the secondary

 ISBN: 978-1-61208-126-7

resource manager when the first manager terminates unpredictably.

ACKNOWLEDGEMENT

This work was supported by the Polish Ministry of Science and Higher Education under grant number 6266/B/T00/2010/39.

REFERENCES

[1] A. F. da Conceicao, J. Li, D. A. Florencio, and F. Kon "Is IEEE 802.11 ready for VoIP," IEEE Workshop on Multimedia Singal Processing, October 2006, pp. 108-113

[2] T. Maseng, "Wireless Tactical Local Area Network," Multinational CDE, Pre-Symposium Workshop, Sundvollen, Oslo, Norway 2002

[3] J. Lopatka and R.Krawczak, "Military Wireless LAN Based on IEEE 802.11b Standard", in Military Communications, Meeting Poceedings RTO-MP-IST-054, Poster 2, France 2006, pp. 21-28

[4] S. Srivathsan, N. Balakrishnan, and S. S. Iyengar, Guide to Wireless Mesh Networks - Scalability in Wireless Mesh Networks,Springer 2009

[5] S. Choi and J. Yu, QoS Provisioning in IEEE 802.11 WLAN, John Wiley & Sons, Inc,2006.

[6] H. Yoon, J. W. Kim, and D. Y. Shin, "Dynamic Admission Control in IEEE 802.11e EDCA-based Wireless Home Network," IEEE Consumer Communication and Networking Conference, January 2006, pp. 1-5

[7] P802.11e – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements, Nov. 2005

[8] W. Wang and S. C. Liew, "Solutions to Performance Problems in VoIP over 802.11 Wireless LAN," Transaction on Vehicular Technology, 54(1), 2005, pp. 336-384

[9] N. Hegde, A. Proutiere, and J. Roberts, "Evaluating the Voice Capacity of 802.11 WLAN under Distributed Control," In Proc of LanMan, 2005, pp. 1-6

[10] L. Gannoune, "A Comparative Study of Dynamic Adaptation Algorithms for Enhanced Services Differentiation in IEEE 802.11 Wireless Ad-Hoc Networks," Proc. of AICT/ICIW, February 2006

[11] L. Cai, Y. Xiao, X. Shen, and L. Cai "VoIP over WLAN: Voice capacity, admission control, QoS and MAC," International Journal of Communication Systems, 2006, (19) pp. 491-508

[12] J. Liu and Z. Niu, "A Dynamic Admission Control Scheme for QoS Supporting in IEEE 802.11e EDCA," Proc. WCNC 2007, pp. 3697-3702

[13] P. Wang, H. Jiang, and W. Zhuang, "A New MAC Scheme Supporting Voice/Data Traffic in Wireless Ad Hoc Networks," IEEE Transactions on Mobile Computing, vol. 7, no. 12, December 2008, pp. 1491-1503

[14] J. Z. Haas and J. Deng, "Dual Busy Tone Multiple Access (DBTMA) - A Multiple Access Control Scheme for Ad Hoc Networks," IEEE Trans. Comm., vol. 50, no.6, June 2002, pp. 975-985,

[15] R. Prasad, M. Murray, C. Dovrolis, and K. Claffy, "Bandwidth Estimation: Metrics, Measurement Techniques, and Tools," IEEE Network, vol. 17, no. 6, pp. 27-35, 2003

[16] G. Chelius and I. G. Lassous, "Bandwidth Estimation for IEEE 802.11-Based Ad Hoc Networks," IEEE Transactions on Mobile Computing, vol. 7, no. 10, 2008, pp. 1228-1241

[17] C. Sarr, C. Chaudet, G. Chelius, and I. G. Lassous, "A node-based available bandwidth evaluation in IEEE 802.11 ad-hoc networks," International Journal of Parallel, Emergent and Distributed Systems, vol. 21, 2006, pp. 423-440

[18] J. Romanik, P. Gajewski, and J. Jarmakiewicz, "Performance enhancement of Wi-Fi ad-hoc network for VoIP support," Proc. of Military Communications and Information Systems Conference, Wroclaw, 2010, pp. 525-535

[19] S. Garg and M. Kappes, "An experimental study of throughput for UDP and VoIP traffic in IEEE 802.11b networks," WCNC, March 2003, pp. 1748-53,

[20] F. Anjum, et al., "Voice performance in WLAN networks - an experimental study," Proc. IEEE Globecom, 2003, pp. 3504-3508

[21] M. Bednarczyk and M. Amanowicz, "Wireless Relay Control Protocol (WRCP)," Proc. of Military Communications and Information Systems Conference, Bonn, Germany, September 2007, ISBN 978-3-934401-16-7

[22] H. Zhang, J. Zhao, and O. Yang, "Adaptive Rate Control for VoIP in Wireless Ad Hoc Networks," Proc. of IEEE International Conference on Communications, 2008, pp. 3166-3170

[23] J. Barcelo, B. Ballalta, and C. Cano, "VoIP Packet Delay in Single-Hop IEEE 802.11 Networks,"WONS 2008, Barcelona, pp. 77-80

[24] S. Sangho and H. Schulzrinne, "Measurement and Analysis of the VoIP Capacity in IEEE 802.11 WLAN," IEEE Transactions on Mobile Computing, vol. 8, 2009, pp. 1265-1279

 ISBN: 978-1-61208-126-7

QoS signaling for Service Delivery in NGN/NGS Context

Noëmie Simoni, Yijun Wu
Network and Computer Science department
Institute Telecom/TELECOM ParisTech
Paris, France
e-mail: simoni@telecom-paristech.fr,
yijun333@hotmail.com

Chunyang Yin
Network and Computer Science department
Institute Telecom/TELECOM ParisTech
Paris, France
e-mail: chunyang.yin@gmail.com

***Abstract*—The evolution of next generation network raises the issue of exploiting customizable services for a nomadic user within the environment of heterogeneity and mobility. In this context, the E2E QoS (Quality of Service) always remains as a challenge. In fact, the E2E supposes that apart the network resource, the service should be taken into account as an important resource, in the form of service component, i.e., we should also provision and monitor the service resource and verify the conformation to the QoS contract. In this paper, we propose a "dynamic user oriented end-to-end QoS signaling" for the service delivery in order to provision and dynamically negotiate the different services in the NGN environment. This end-to-end signaling extends to process the service delivery to maintain QoS continuity. The related sequence diagrams are defined and a demonstration is implemented with the purpose of presenting the feasibility of our proposals.**

Keywords-NGN; service delivery; QoS signaling

I. INTRODUCTION

As networks are evolving towards NGN (Next Generation Network), the environment is becoming more and more complex. NGN is characterized by heterogeneity mobility and user centric.

The deployment of such NGN raises the issue of development of Next Generation Service (NGS). The mobility challenges to offer service continuity in a personalized way for end users. Moreover, the personalization puts the user in the centre position in the architecture (user centric). An end-to-end service session should integrate user's preferences with the dynamic context, which includes end user terminals, access networks, core networks and services (heterogeneity).

If we follow the same strategy, then the user becomes the center of the consideration. Therefore, first, since the user is probably across the heterogeneous environment, to deal with the different providers or to gain the benefits from this evolving environment. For example, a travelling user is in a café bar where the WiFi service is offered. He is glad to grasp this chance to check his emails, but he has only the email account settings on his laptop and he's not in the mood to configure his telephone for that moment. Therefore, the different terminals are desired to be integrated before the execution phase of a user's service session. On the other side, user is also willing to benefit from the NGN mobility supports to have the service continuity. Mobility allows the end user to communicate regardless of location, used device, access mode and multiple spatial network domains. Finally, he wants to always keep his personality when the system choosing or adapting the services for him through his pertinent preferences.

All these requirements lead us to focus on the service delivery with the service continuity. The service delivery we mentioned above is different with the traditional media delivery. Media delivery focuses on the data transfer solution treating the service continuity by shifting the access point and the corresponding supporting services in the network layer. The QoS is obliged to be recalculated and sometimes the delivery is interrupted.

What's more, to insure the E2E (End-to-End) QoS, today's media delivery E2E QoS solution, aiming the network resources, is no more sufficient. The service should also be taken into account as an important type of resource as well as network resource. To compose the service as service components is a possible way to separate the service resources from the network resources. We should not only provision and monitor the service resources as well as the network resources, but also dynamically verify the conformation to the QoS contract established between service providers and user.

In fact, the aiming user-centric session should allow users to access services in a customized way (composition of service) with QoS continuity. Therefore, a more flexible E2E QoS signaling is desired to support the service delivery and guarantee the QoS continuity. Its principle task is to negotiate the QoS and user information among service components for subscribers across any mobile or fixed network with any user appreciated equipment during the provisioning phase. During the real time of the user session, it can help to dynamically manage the session on the QoS level to have the QoS continuity.

In this paper, we propose a QoS signaling which covers the end-to-end provisioning for demanded service and maintains the provided service conform to the SLA (Service Level Agreement) during the real time service session. Quality of Service in this paper is seen from end-user's standpoint which means that the QoS assessment should be performed regards to the users' requirements.

The remainder of this paper is organized as follows. In Section II, the requirements of user-centric session are introduced. In Section III, the related works in signaling protocol are presented and analyzed. In Section IV, a dynamic E2E QoS signaling is proposed based on a service

 ISBN: 978-1-61208-126-7

and sub-network integrated architecture. In order to examine the feasibility of our proposition, in Section V, the functional specification is detailed by explaining a user case and scenarios which take user's needs into account with corresponding sequence diagram, and the implementation part is presented. Finally, Section VI is the conclusion and gives out the perspectives.

II. USER CENTRIC SESSION REQUIREMENTS

We have mentioned that there's an important characteristics of NGN: user centric. However, what will be introduced by this change? The telecommunication world evolves and becomes now user centric in opposition to system centric (behavior is constrained by the system) and network centric (behavior is constrained by the network). Therefore, user-centric requirements are expressed by user-related information, such as QoS parameters and user's preferences etc. This information can be defined in the user's profile. On the provider side, QoS commitments should match these user's requirements. These commitments can be defined in the customer contract SLA. As Figure 1 shows, the user's profile in the centre represents the user in the E2E session. This conception contains the essential requirements as below:

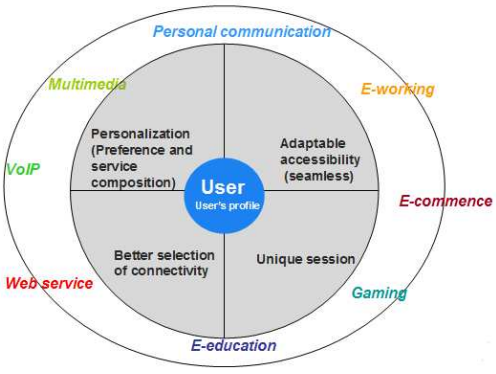


Figure 1. User-centric session requirements

- Personalization according to user's preferences and the user profile. In order to meet the personalization of service delivery, a more flexible deployment of services (service aggregation) as well as a loose coupling architecture between service elements (service composition) is appreciated.
- Unique user session for converged services delivery from different providers in a top-down way. It means that we don't need lance several sessions for different media service demanded by user. It is possible to have different media demanded in one session through difference transactions.
- Better selection of connectivity with QoS in a dynamic manner according to user's preferences (QoS, location, agenda, etc.) for the media delivery.
- Adaptable accessibility to services regardless of terminal used. End users may have different user equipment (UE) in a Seamless Userware vision according to user's preferences [1].

III. RELATED WORK

An end-to-end session goes through the user's equipment, the connectivity network session (access and core network) and the service session. In each sub-session, it has his own method for the QoS issue, for example, GPRS (General Packet Radio Service) Tunnelling Protocol (GTP) used in between SGSN (Serving GPRS Support Node) and GGSN (Gateway GPRS Support Node) allows the SGSN to adjust QoS parameters in a session GPRS on a user's behalf (PDP context); Session Initial Protocol (SIP) used between user and application server works with Session Description Protocol (SDP) for defining and negotiating the QoS parameters of media streams. We find that there is nothing for the service level.

For the requirements of user-centric session, signaling should support broadening personalization coverage with QoS through composition of services in service levels to enhance capability of adaptation in mobility and heterogeneous environment. In this new context, we have evolved from a client/server architecture for a demanded application to a distribute architecture with a composition of services for a continuity of session demanded.

Before presenting our proposals, we analyze the existing related work on the architectural proposals (IMS) and signaling proposals (SIP, NSIS) aiming at end-to-end QoS in this section. IMS (IP Multimedia Subsystem) architecture gives a converged network control layer. SIP is a signaling protocol of the session layer for media QoS control with the media. NSIS (Next Step In Signaling) is a signaling protocol of network layer to transferring with QoS parameters and. These proposals are in the network level to maintain the QoS, and there is nothing for the dynamic resource reservation of service.

A. IMS

IMS architecture defined by 3GPP (3rd Generation Partnership Project) creates one common access independent signaling platform for providing multiple services [2][3][4]. It implements the QoS mechanisms in management and control planes, i.e. policy control in the management plane and admission control in the control plane. In the TISPAN (Telecommunications and Internet converged Services and Protocols for Advanced Networking) NGN architecture[5][6][7], not only core networks but also access networks or even terminals on the user side have QoS related control and management functions.

IMS integrates the network convergence but not the service convergence. In addition, IMS provides access to a service while still following client-server architecture as a tight coupling. In fact, we need a loose coupling for the service trans-organization in order to favor the personalization of service.

 ISBN: 978-1-61208-126-7

B. SIP

SIP as the basic protocol at IMS control layer and one of the multimedia communication system framework protocols [8], is the application layer protocol used to establish, change, or terminate a multimedia session. SIP enables flexible interaction of several media in one session.

As far as QoS is concerned, SIP uses SDP to describe the media in the session and negotiate QoS requested in the network layer [9]. Moreover, SIP can filter information according to User Profile to implement application servers before establishment of a session.

However, SIP does not cover description of service component behavior, and is not even able to communicate the QoS related information among the components in order to re-provision services during an active session in a mobility environment committed to user's contract.

C. NSIS

NSIS focuses on developing a protocol to manipulate QoS resource states along the data path in the network. NSIS is concerning to media delivery. But their work does not cover all the service resources in order to cover overall service delivery which is more direct user relationship.

The QoS NSLP (NSIS Signaling Layer Protocol) proposed by NSIS (Next Step in Signaling) work group in IETF (Internet Engineering Task Force) provides flexibility on patterns of signaling messages that are exchanged [10]. The proposed specification of QoS (QSPEC) carries a collection of objects that can describe QoS specifications in a number of different ways, named QoS Desired, QoS Available, QoS Reserved and Minimum QoS. A generic template which contains object formats for the QoS description has been designed to ensure interoperability while using the basic set of objects [11]. NSIS focuses on developing a protocol to manipulate QoS resource states along the data path in the network. NSIS is concerning to media delivery.

However their work does not cover service resources which are essential in service delivery, where we should also reserve the QoS for each service component in a session.

Facing to the NGN/NGS needs, the desired solutions should concern the QoS control for all the resources in a user-centric session, not just the network ones but also the service ones. The related work analysed above is still just in the network layer for the QoS control of media delivery.

IV. PROPOSITION: USER-CENTRIC QOS SINGALING

From a user centric point of view, the E2E QoS could be enhanced by including service components as resources. In order to satisfy the E2E QoS in the user-centric session, we propose a "dynamic user oriented E2E QoS control" supporting QoS provision and dynamic management (negotiation and adaptation). We develop this signaling on the existing SIP protocol framework and call it enhanced SIP (SIP+). SIP+ is able to circulate the media description and service component description for service delivery provisioning. We introduce firstly the VSPN (Virtual Private Service Network) architecture (§A) on which our proposition is built, then we explain the E2E QoS provisioning protocol (§B). Finally, we detail the dynamic QoS signaling which converges the management information in control messages during the exploitation with a state chart (§C).

A. Architecture dimension

In order to manage the service resources, we constitute VPSN with networked Service Elements (SEs) (Figure 2). This network is virtualized because the service components have sufficient abstraction features and they are mutually sharable. This network is private because it responses to a service request of a particular customer with specific QoS needs. The VPSN translates the logic of the requested service and links the service nodes according to "semantic routing" for the service composition. The VPSN proactively maintains this QoS to meet the customer's SLA. Concerning the QoS communication and interworking solution between the separated service and network layer, based on IMS platform, in [12], authors have proposed a binding mechanism in the E2E user-centric session to correlate the information generated in the three parts (User equipment, network and service). These three levels would be aware of and self-adapt to any change for the dynamic QoS information management. This information are stocked in database on two sides: Infosphere, in user side, which provides personalization and ambient resources information around end-user; and an enhanced UPSF (User Profile Server Function) in operator's side which offers an informational inference on all the resources [13]. The link is thus established integrating with sub-networks.

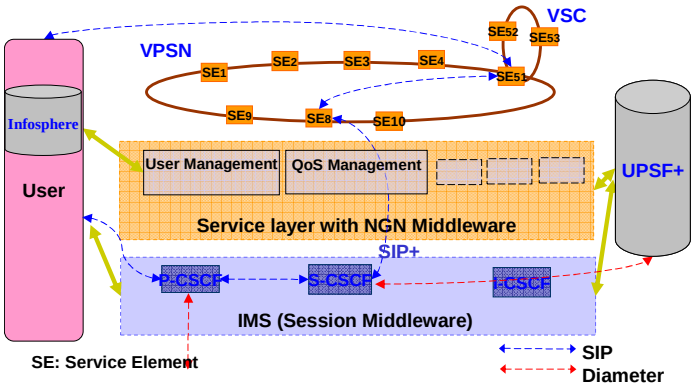


Figure 2. Global architecture of SIP+

We propose a VSC (virtual service community) to manage the QoS and functional equivalent service components in a community. VSC is a group of SEs that has the same functionality with an equivalent QoS. In case that a component is no longer suitable for a user session, another member could replace it and keep the QoS continuity.

We map the proposal on the ETSI (European Telecommunications Standards Institute) IMS. Service components are in the different container (for example: Application Server). The NGN middleware offers the basic services, which enables QoS management, user management

 ISBN: 978-1-61208-126-7

and other management functions. Service components are thus managed by NGN middleware. The NGN middleware selects and invokes services at the reception of SIP message from the IMS session middleware. These service components are chained into VPSN according to the service logic by the SIP+. Moreover, the service components can negotiate QoS information with end users through SIP+.

B. E2E QoS provisioning protocol

Enhanced SIP works in service layer to provision the resources for the E2E user-centric session based on our QoS model (four criteria: availability, delay, reliability and capability). This unified QoS model is applicable to all the resources in a session. These criteria can be applied to any QoS classification (Diffserv, Interserv, etc.) and can be also easily measured according to specific parameters. During the service's deployment and provisioning, these four criteria are divided into three categories: conception value, current value and threshold value.

The **conception value** is decided at the phase of service conception. It introduces the maximum possibilities of the node's treatments and the link's interactions.

The **current value** is calculated during provisioning and exploitation to reflect the service's behavior in real time.

The **(minimum/maximum) threshold value** defines the range on which the node normally operates. They are alert thresholds avoiding the current value exceeds the limitation.

Under the help of these modeling tools, the SIP+ could provide a QoS description in high level covering an overall service requested by end-user in a top-down declination (service node QoS, network QoS, equipment QoS) approach, as Figure 3 shows.

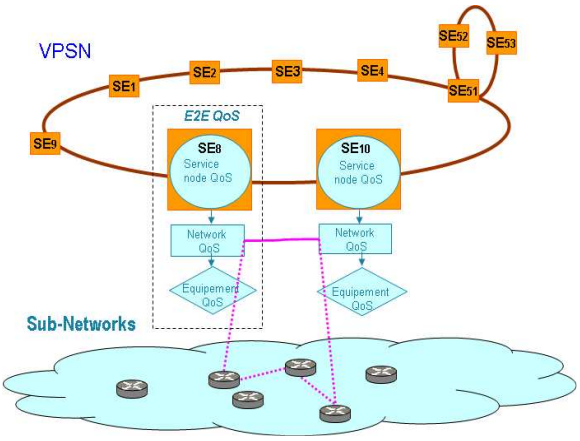


Figure 3. E2E QoS description

The service node QoS contains the non functional characteristics of service. The network QoS collects the routing table which records the QoS of all the possible paths in the transport layer. The equipment QoS is the QoS view of the machine (CPU and memory etc).

The recursive calculation of the QoS between the levels enables the E2E QoS aggregation across multiple providers. If service nodes are activated, the real-time QoS condition is recalculated and updated in the QoS table. The latter will probably cause the recalculation of the network QoS.

NSIS has four kinds of QSPEC (QoS Specification) objects which consist of a number of parameters describing the condition and constraints of traffic as well as traffic classifier for the resource reservation in the network layer. The SIP+ aims at provisioning service resources in the same way as NSIS. Therefore, we propose a QoS description template in the SIP+ which has two objects: Demanded QoS and Current QoS (Figure 4). In each object, QoS parameters are classed into four QoS criteria.

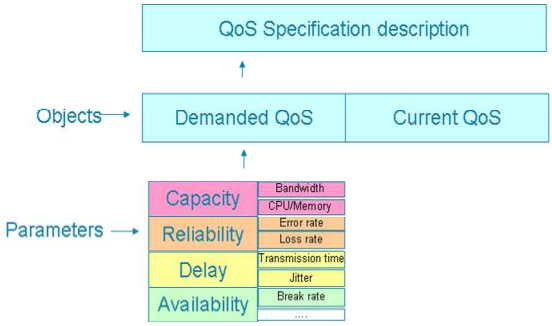


Figure 4. SIP+ QoS description template

In each node of VPSN, it has a QoS agent which stores a contracted QoS (Threshold range values) according to the SLA. The Demanded QoS and Offered QoS are negotiated during the phase of provisioning after service deployment according to the Contracted QoS. These QoS parameters are taken into account during the QoS provisioning. Thus, the service components have knowledge of the contract to fulfill and the image of its current performance.

The service request (INVITE) message contains the demanded QoS. If the demanded QoS value is less than the current value; the demanded SE gives an "OK" response with the offered QoS. This service component will be activated. On the contrary, if the desired QoS value is more than the current value, the service component will not be activated. When the E2E provisioning ends, the session is activated. Figure 5 shows the corresponding QoS values in existing NSIS, SIP and our proposal SIP+.

NSIS (Network)	SIP (Network)	SIP+ (Service)
QoS Desired	Desired	Demanded QoS (conform with contracted QoS)
QoS Available	Current	Current value (Monitoring Value)
QoS Reserved		
Minimum QoS		

Figure 5. SIP+ vs. SIP and NSIS

 ISBN: 978-1-61208-126-7

C. *Dynamic QoS control for service delivery*

Dynamic QoS signaling gives the possibility of corrective actions to various problems (mobility, user's preference) during the service delivery in order to maintain the QoS (Session continuity). New service components could be added; meanwhile activated service components could be replaced by others. Service component's QoS condition (In contract/out contract) in the management system, which is got from comparison of current value and threshold range values, could be notified from time to time during the service delivery via the SIP+ NOTIFY message so as to maintain dynamically the QoS conform to the contract. The state chart of QoS control procedure is shown in Figure 6.

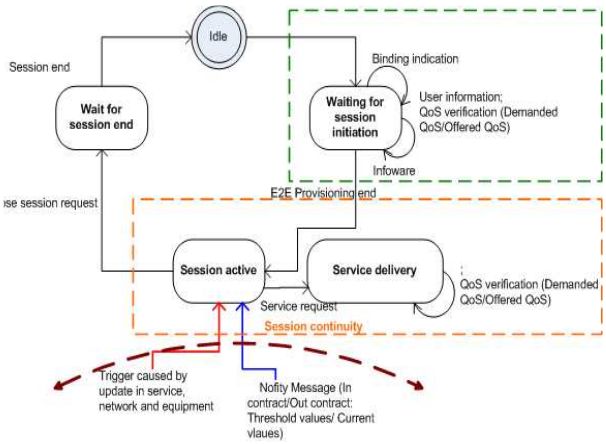


Figure 6. State chart of QoS control in service layer

In an open session, when the QoS condition in one node changes, i.e., the current value exceed the threshold value, the VSC to which this service component belongs will firstly perform self-management to find another service component with the equivalent QoS to replace it. VSC manages the SEs in terms of QoS management. For instance, VSC could look for the suitable SE that can replace the QoS degraded one. If it fails, the QoS control could interact with the user system in the database and re-provision the QoS in the VPSN in order to find another node with a QoS commitment according to user's preference (Figure 7). Meanwhile, the sub-network establishes a QoS path simultaneously. Therefore, the user-centric session keeps continuous.

In the end-to-end session QoS signaling, we define four management states for each service node. The state chart of service component management is shown in Figure 7.

UNAVAILABLE means the node is not accessible for the user or is unreachable due to user's mobility.

AVAILABLE means the node can be accessed but not activated yet.

ACTIVABLE means the node is ready to use. The node is considered to be activated when it has been chosen to be a part of the VPSN and can be executed at any time.

ACTIVATED means the node joins a real time transaction of a user centric session; its resource is being consumed. In this state, the QoS control agent checks the resource in real-time. If its behavior conforms to QoS contract, the state is changed to "IN CONTRACT". If the behavior does not conform to QoS contract, the state is changed to "OUT contract". And the VSC launches the process of self-management to find another service component to replace it.

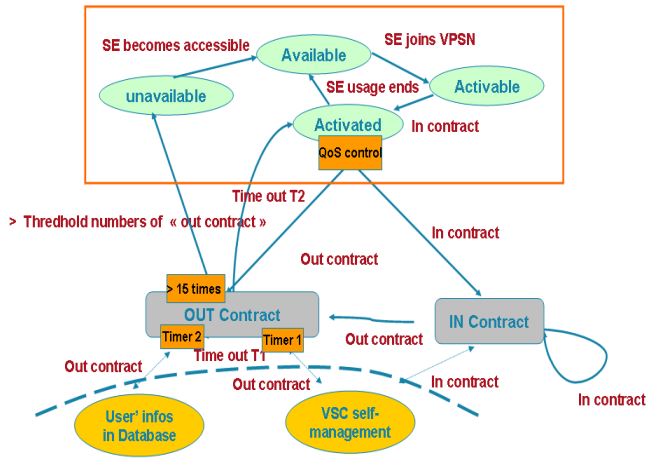


Figure 7. State chart of service component management

IN CONTRACT means that the QoS condition is in the scope of contract signed by the user and operator.

OUT CONTRACT means that the QoS condition is out of scope of contract signed by the user and operator. After receiving the first "OUT contract" in the message, the node arms a timer (Timer 1) to wait for the VSC processing. If the VSC does not find a solution till timeout (Time 1), the node will solicit database (inforsphere & infoware) to modify the contracted QoS according to the user preference. The node arms a Timer 2 for this process. Until Timer 2 time out, the QoS control agent checks the current resource with the updated QoS contract in ACTIVATED.

If the activated node receives continuously the "Out contract" in the header of NOTIFY messages that the received number exceed the threshold numbers for deactivating, the node turns into **UNAVAILABLE** state.

In addition, we identify the events that cause transition of the state. User Initiated Events (User preference), and Service Initiated Events (self-management of service components in the virtual service community) are external events that trigger state change of service node. Meanwhile the QoS condition (In contract/Out contract) and the timer in the entities are needed to be notified in the VPSN during the service delivery. The latter are therefore identified as internal events.

V. FUNCTIONAL SPECIFICATION AND DEMONSTRATION

In this section, we describe a user case and related sequence diagrams with the purpose of presenting the functional specification of the concepts proposed and a demonstration. Based on a user case (§A) highlighting the specificities of the NGN context and the diagram sequences which are relevant to the scenario in our user case, we will present finally the demonstration (§B).

 ISBN: 978-1-61208-126-7

A. *User case*

The user case (Figure 8) clearly highlights the specificities of the NGN context. It shows terminal mobility, user mobility and session mobility with QoS continuity while the user moves across heterogeneous networks according to his preferences. In an "E2E user-centric session" vision, user could enjoy different types of media in one unique session in a customized way. The scenarios details are described as below.

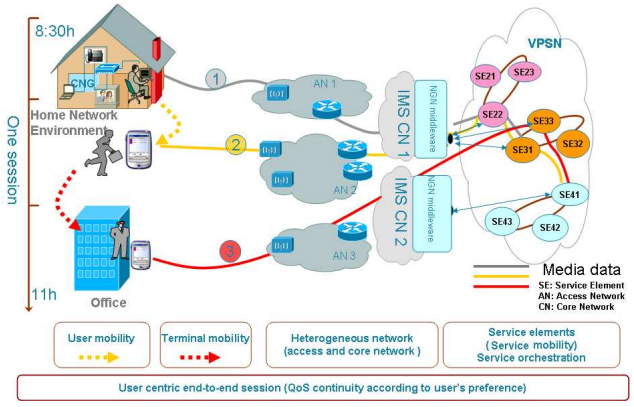


Figure 8. User case

While at home, the end-user Bob starts his PC at 8h30 in the morning and engages in a Video service SE22 and Web service SE 31. Bob uses his home network environment and accesses to the service through an Access Network AN1. Before leaving his home, Bob switches terminals (User Mobility), leaving his PC for his PDA. At this point, Bob, as a pedestrian, prefers to receive his messages in vocal mode rather than in text mode (User's preference). Therefore, a Text to Voice service component (SE41) is invited into the opened session to adapt the new terminal according to user's preference. When arriving at work, Bob's PDA is still attached to the same service but the Web service element (SE31) is replaced by the SE33 in the same VSC (service mobility) for keeping QoS continuity in one session. At 11AM, Tom closes the session, which he opened at home at 8h30AM.

For the user, all the operations are transparent.

In our demonstration, we will implement a scenario that the SE31 is replaced by SE33 with QoS continuity. We call this procedure as "service mobility". The SE31 (like other SEs in the VPSN) notifies its QoS notification (IN/OUT contract) to elements of its VSC in a peer to peer way. When SE31 is "OUT Contract", the VSC finds another SE (SE33) to take place of SE31 within a limited time. The SE33 is then introduced in this session. In this way, the service components in one session can always keep their QoS "IN contract" according to SLA, the service mobility is done seamlessly.

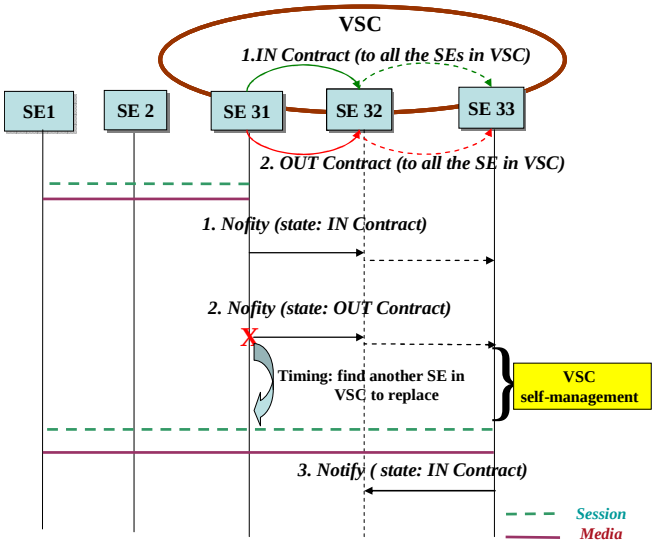


Figure 9. Service mobility related sequence diagram

The management result "IN" or "OUT" Contract is delivered in the signaling message SIP NOTIFY message from time to time in the VPSN.

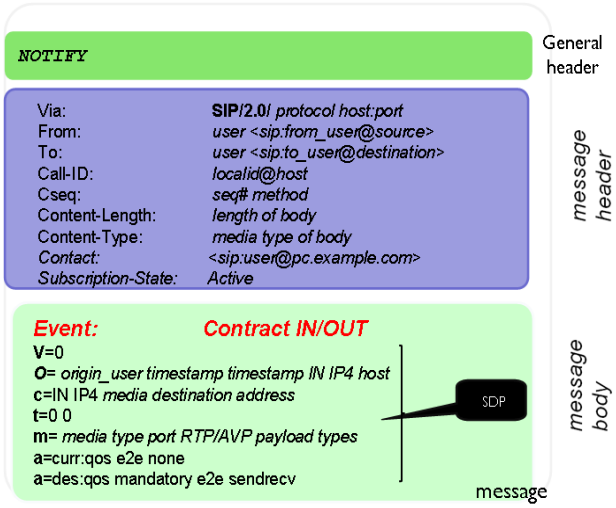


Figure 10. Structure of NOTIFY message

The structure of SIP NOTIFY message is shown in Figure 10. A SIP message, both the request and the response, contains a start-line followed by one or more headers and a message body. The CONTRACT IN/OUT is proposed to add into the message body (in red) of SIP to inform the management information during the process of service delivery. The NOTIFY body is used to report the condition on the resource being monitored (current value comparing with threshold values).

B. *Specification of experimental platform and demontration*

The identified scenarios are experimented on our

 ISBN: 978-1-61208-126-7

platform [14] in laboratory. This platform is consisted of the OpenIMS Module (FOCUS) [15] which contains the essential blocks of IMS core (P-CSCF, I-CSCF, S-CSCF) and HSS+ in Oracle [16], and service platform (Service is offered as a composition of service elements considering the QoS management, security, mobility management and charging aspects) developed in SUN's GlassFish [17] container with SIP JAIN supporting SIP server development (for SIP+ module). The detailed architecture of our technical platform is shown in Figure 11.

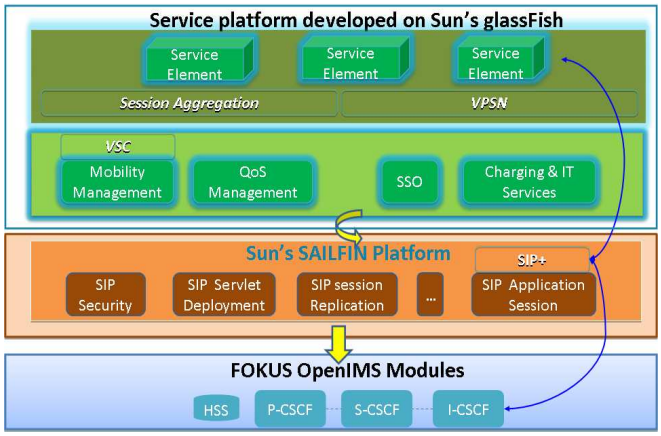


Figure 11. Experiment platform

SIP+ module developed basing on SIP JAIN, notifies the QoS states among the service elements, and relays QoS information between service layer and IMS control layer. In our demonstration, we present the QoS management information (In/Out contract) exchanged between two service elements in VSC. As the Figure 9 shows, the SE31 is active in the VPSN, and it sends the NOTIFY messages to the SE32 which is in the same VSC for informing the QoS event (In/Out contract). The global architecture of implementation is shown in Figure 12. The Data Base enables the SIP server to compare the current QoS value with the threshold QoS value. The result will tell if the SE's QoS is in or out contract.

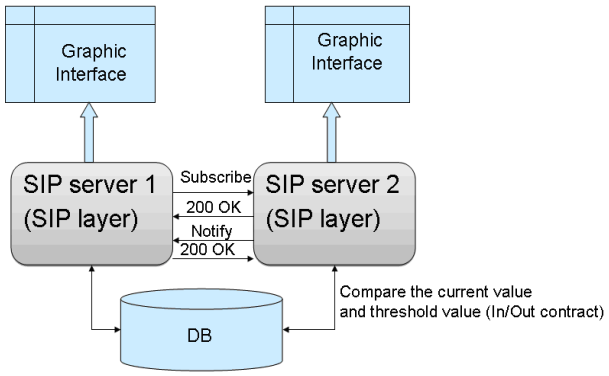


Figure 12. Global architeure of implementation

In order to easily show the messages exchanged between two SEs, we create a graphic interface for each service element. It is shown in Figure 13.

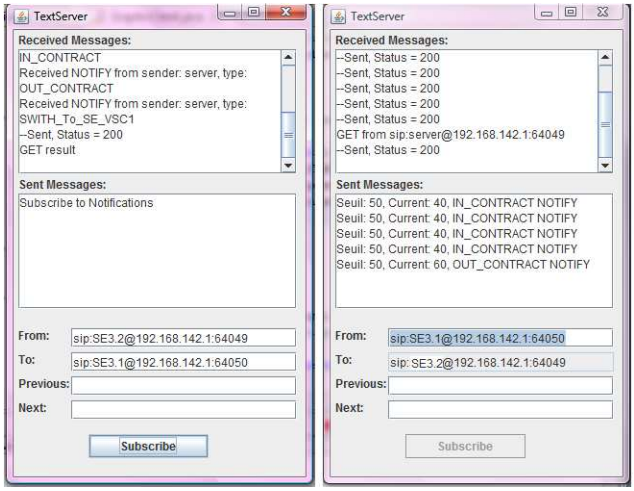


Figure 13. Messages exchanged in graphic interfaces

VI. CONCLUSION AND FUTURE WORK

End users expect to have a continuous comprehensive service throughout the whole session while moving (terminal mobility) or changing terminal (user mobility). During this session, a service is considered as a composition of elements in order to adapt to any change (session mobility). In such a user-centric approach, today's control and management approaches aim at how to managing the QoS through media delivery. However, those solutions are unable to cover this E2E QoS issue to allow user-centric oriented service to be carried out properly.

In this paper, we propose a QoS signaling (SIP+) in the service layer which covers end-to-end provisioning and SLA contract conformity. During the provisioning, SIP+ takes the behavior of the service component and the user's expectation into consideration as well as media resources in the network. During the service delivery, information from management system (monitoring results "In contract" and "Out contract") can be notified in the service layer via SIP+ control message to maintain E2E SLA. Finally, a user case and scenarios with related sequence diagrams are specified for analyzing the feasibility of our proposal. And a scenario is experimented in our platform. The following work is to integrate our implementation within our global platform (VirtuOR's virtual router [14], FOCUS OpenIMSCore [15] and SUN's GlassFish [16] [17]) for analyzing the performance.

 ISBN: 978-1-61208-126-7

ACKNOWLEDGMENT

The authors thank to all the members of ETSI STF 360, and also thank to the members of ANR's project UBIS for their work on our platform.

REFERENCES

[1] Sasan Rostambeik, Noëmie Simoni, Antoine Boutignon: Userware: A framework for next generation personalized services. Computer Communications 30(3): 619-629, 2007

[2] Miika Poikselkä, Georg Mayer, Hisham Khartabil, and Aki Niemi, "The IMS: IP Multimedia Concepts and Services", 2nd edition, Wiley, 2006

[3] 3GPP TS 23.228 version 8.6.0 (2010-06), Universal Mobile Telecommunications System (UMTS); LTE;IP Multimedia Subsystem (IMS), Stage 2, Release 8

[4] ETSI TS 123 228 V8.6.0 (2008-10): Digital cellular telecommunications system (Phase 2+)

[5] ETSI TS 185 006 V2.0.0 (2008-03), Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Customer Devices architecture and Reference Points

[6] ETSI ES 282 001 V2.0.0 (2008-03), NGN Functional Architecture

[7] ETSI TS 185 005 V2.0.0 (2007-07), Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Services requirements and capabilities for customer networks connected to TISPAN NGN

[8] RFC 3261: SIP (Session Initiation Protocol) July, 2002-06

[9] 3GPP TS 24 229 V10.1.0 (2010-09): 3rd Generation Partnership Project; Technical Specification Group Core Network; IP Multimedia Call Control Protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP) Stage 3, Release 10

[10] NLSP for QoS. Available: http://www.ietf.org/html.charters/nsis-charter.html , December, 2010

[11] QoS NSLP QSPEC Template. Available: http://www.ietf.org/html.charters/nsis-charter.html

[12] N. Simoni and Y. Wu.: NGN Charging in enhanced IMS. ICN'08: 87-92, Cancun, Mexico

[13] N. Simoni, C. Yin, and G. Du Chéné.: An intelligent user centric middleware for NGN: Infosphere and AmbientGrid. COMSWARE 08: 599-606. Bangalore, India

[14] http://www.virtuor.fr/, December, 2010

[15] http://www.openimscore.org/, January, 2011

[16] http://www.oracle.com/index.html, Januarry, 2011

[17] http://glassfish.java.net/public/downloadsindex.html#top, December, 2010

 ISBN: 978-1-61208-126-7

Impact of Bit-Flip Combinations on Successive Soft Input Decoding of Reed Solomon Codes

Obaid ur Rehman and Natasa Zivic
Institute for Data Communications Systems
University of Siegen
Hoelderlinstrasse 3, 57076 Siegen, Germany
email: obaid.ur-rehman@uni-siegen.de and natasa.zivic@uni-siegen.de

***Abstract*- Soft Input Decoding of Reed Solomon codes using successive iterative bit-flipping was introduced earlier. The proposed method uses a concatenation of Convolutional and Reed Solomon Codes. It was shown that the proposed method not only improves the coding gain, but also helps in circumventing the miscorrections, that occur with the given probability, using the hard decision Reed Solomon decoder. In this paper, the impact of bit-flip combinations on the soft decision Reed Solomon decoder is analyzed. Increase in the bit-flip combinations results in an increase in the error locator set of the iterative decoding algorithm. It is shown that by increasing the possible bit-flip combinations using a threshold, the coding gain of the concatenated codes increases while the miscorrections rate decreases. The impact of the increase in the threshold on the decoder performance is also investigated in this paper. Simulations are performed for different code rates and an improvement in the coding gain is obvious through the simulation results presented. Moreover, the simulation statistics show a decrease in the miscorrections with an increase in the size of the bit-flip combinations.**

Keywords- Concatenated Codes; Reed Solomon Codes; BCJR; Successive Decoding; Miscorrection Detection.

I. INTRODUCTION

An iterative method for the Soft Input Decoding of Reed Solomon (RS) codes with successive iterative bit-flip decoding was shown earlier [1]. The proposed method considered concatenated Convolutional/RS Codes with CRC for error detection. Code concatenation is a method to increase the decoding capability of the individual codes by concatenating them together in a serial or parallel manner. Code Concatenation was first introduced by G. David Forney [2]. It is a method of combining two (normally different) codes to obtain a better Bit Error Rate (BER). They have many practical applications such as in the Compact Disc technology, Digital Video Broadcast, WiMax and the Voyager program. A typical concatenated arrangement is to use the hard decision Reed Solomon codes as the outer code and the soft decision Viterbi code as the inner code with an interleaver in between to break the burst of errors.

In this work, concatenated codes with Reed Solomon code as the outer and a convolutional code with the corresponding BCJR/MAP [3] decoder as the inner code are considered. The Maximum a-posteriori Probability (MAP) decoder outputs the hard decoded data (the binary form of decoded data) along with the Log Likelihood Ratio (LLR) or the L-Value of each and every decoded bit. These LLRs (or L-Values) give an estimate of the confidence on the hard-decision bits. Thus lower the absolute value of the LLR (i.e., |LLR|) of a bit, the least reliable it is and vice versa. The hard decision RS decoding is done on the output of the MAP. The hard decision RS decoder will result in either a decoding success or a decoding failure. There is, however, a known probability of decoding error [4]. The decoding error is also termed as "miscorrection" in literature. Decoding error occurs when the decoder decodes the received word to a codeword other than the transmitted one. RS decoder "sees" the decoding error as successful decoding, so in this work a CRC code is used to detect and then avoid the decoding errors, with a high probability.

In case of a decoding failure and/or error, a successive iterative algebraic decoding on the Reed Solomon codes is performed with a different combination of bit-flips in each iteration. The number of the bit-flip combinations dictates the size of error locator set used in the successive iterative decoding. The size of the error locator set is directly proportional to the performance and the computational costs. In this work, the impact of an increase in the bit-flip combinations on the decoder, introduced in [1], is studied both analytically and via simulations.

It is shown that increasing the bit-flip combinations will increase the coding gain. The number of iterations, however, doubles with every next bit added to the bit-flip combination set. The idea of inversion of the least reliable bits was introduced earlier in Chase decoding algorithms [5] and the Generalized Minimum Distance (GMD) decoding [6] algorithms. Recently, the idea of bit-flipped decoding has developed much interest, mainly, due to the rediscovery of Low Density Parity Check Codes [7]. Similarly, soft decoding techniques for the RS codes were initiated in the pioneering work done by [8] and [9]. More recently, the author in [10] has treated the subject in more detail by introducing some soft decoding techniques using bit-level soft information.

The rest of this paper is organized as follows. In Section II, the Soft Input Decoding of Reed Solomon Codes with miscorrection detection and avoidance [2] is discussed briefly. In Section III, the impact of bit-flip combinations on the

 ISBN: 978-1-61208-126-7

scheme is discussed. In Section IV, the simulation results are presented with different bit-flip combinations and different code rates. Finally, a conclusion is given in Section V.

II. SOFT DECODING OF REED SOLOMON CODES WITH MSCORRECTION DETECTION AND AVOIDANCE

In this section, the encoder and decoder for successive iterative decoding of RS codes with miscorrection detection and avoidance is introduced briefly.

A. The Encoding Process

The Encoding process used in this study is depicted in Fig. 1. Here, the encoder is a simple concatenated RS-Convolutional encoder. This code concatenation is widely used in many popular protocols such as the Digital Video Broadcast (DVB-S) and WiMax.

CRC of m-bits is computed on the data of length k-1 and the data along with the appended CRC are then encoded with a systematic RS encoder. In this case, the RS encoder produces a codeword (c) of n-symbols (each of m-bits over GF(q)), where $q=2^m$ and n = q-1 (or 2^m-1). The codeword c is represented as,

$$c = (c_1,c_2,c_3,\ldots\ldots\ldots,c_n), \qquad c_i \in GF(q) \qquad (1)$$

This codeword (c) is then converted into its binary equivalent for encoding with a convolutional encoder. The binary equivalent of "c" can be formally expressed as,

$$c_b=(c_{1,1},c_{1,2},\ldots.,c_{1,m},c_{2,1},c_{2,2},\ldots..,c_{2,m},\ldots\ldots\ldots..,c_{n,1},c_{n,2},\ldots\ldots\ldots\ldots\ldots\ldots,c_{n,m}), \qquad c_{i,j} \in GF(2) \qquad (2)$$

This binary-vector c_b is then encoded by the convolutional encoder to produce the message u for transmission. The message u is then BPSK modulated and transmitted over the Additive White Gaussian Noise (AWGN) channel.

The role of the CRC code is to detect the decoding errors [4] made by the RS decoder. This detection helps in avoiding these decoding errors through a successive iterative decoding process.

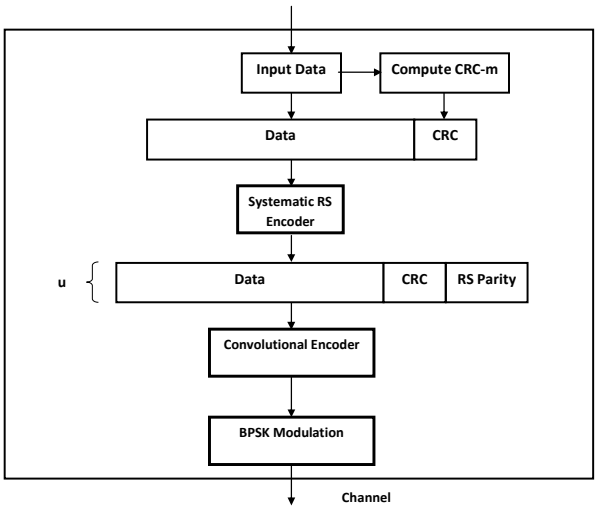


Figure 1. The Encoder

If the received word (r=u+e, where e is the error pattern) has a smaller Hamming distance to another valid codeword, the RS decoder will "miscorrect" it to this closest codeword. Such a miscorrection is called a decoding error [3] and in the decoder introduced in this work a CRC code is used to detect such a decoding error. CRC gives the ability of detecting the decoding errors at the cost of additional redundancy, thereby effectively reducing the code rate. This is catered-for in the simulation results by making comparisons with the corresponding low rate encoders (for fairness). In this work, CRC codes of different lengths for different RS code rates are considered. The CRC code, however, always occupy one symbol in the codeword in order to have a minimum impact on the overall code rate, as discussed in the section on simulation results. CRC is placed in the last m-bits of the RS data part, so that it gets protected from channel noise by the RS parity, allowing for the detection of RS decoder errors

B. The Soft Input Decoder

The successive iterative soft input decoding of the concatenated Convolutional/RS codes is illustrated in Fig. 2. The decoding starts with the Soft Input Soft Output (SISO) Log MAP decoding. The output of the Log MAP decoder is the soft estimate of the decoded data along with their reliability values (so called L-Values or the Log Likelihood Ratios (LLRs)). Hard decisions are made on the output of the MAP decoder and the binary data is fed to the hard decision RS decoder. The RS decoder will either successfully decode the data make a decoding error or fail to decode it, resulting in a decoding failure. These situations are summarized by the three cases given below,

1. The RS decoder succeeds to decode the given word. The CRC is recomputed on the k-1 decoded data symbols and compared with the CRC present in the k^{th} data symbol.
   a. If both the CRCs match, the decoding is successful and the iterative process stops.
   b. If the CRCs do not match, a decoding error is made by the RS decoder and some predefined number of attempt are made to correct the "miscorrected codeword".
2. The RS decoder fails to decode the given word. This happens when the RS decoder is neither successful in rightly decoding the given word, nor in miscorrecting it. In this case the "erroneous word" is not discarded, rather subjected to the iterative bit-flipping technique in order to try and recover the transmitted codeword (c).

In cases 1b and 2, the successive bit-flip decoding of RS codes is performed iteratively with a different combination of bit-flips in each iteration. The bit-flipping is done on the basis of |LLR| values obtained from the MAP decoder. A combination of bits from this "wrongly decoded" codeword is flipped followed by the hard decision RS decoding and CRC comparison. If the CRCs match, there is a success and the successive iterative decoding stops. If not, then the next combination of bits is flipped from the original "erroneous word".

 ISBN: 978-1-61208-126-7

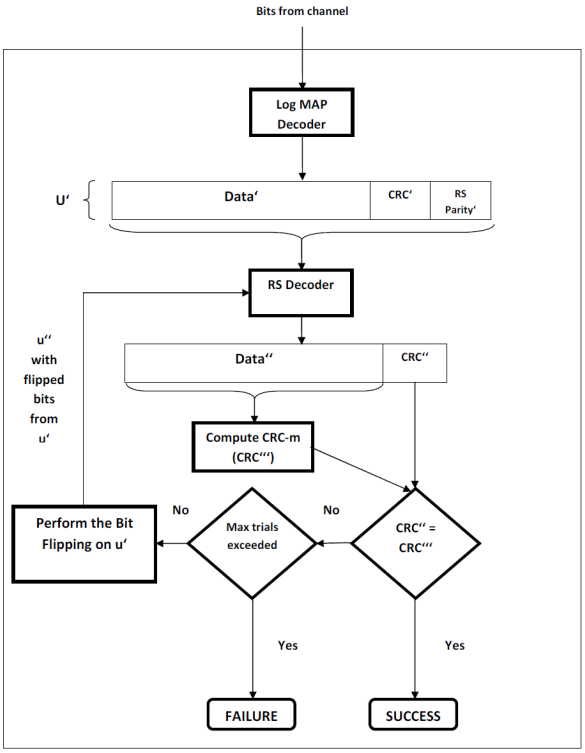


Figure 2. The soft input decoding of RS codes based on Bit-Flipping [1]

This process of bit-flipping and CRC comparison is repeated until either an RS decoding success followed by CRC comparison success or a threshold number of iterations. When the threshold is reached, the decoding fails.

The combination of the bits to be flipped is decided as follows. In the first iteration the least reliable bit is flipped. In the second iteration, the second least reliable bit is flipped. In the third iteration the first and the second least reliable bits are flipped and so on. This is limited by the number of bit-flip combinations chosen in advance, e.g., for 8-bit bit-flip combination a total of 2^8-1 combinations will be considered. The decoding error of the successive iterative bit-flip decoder is bounded by the following equation (for explanation and derivation of the following please refer to [1]),

$$P_E = P_{ERS} * P_{ECRC} \quad (3)$$

where P_{ERS} is the probability of decoding error with the hard decision RS decoder and P_{ECRC} is the probability that the CRC collision will go undetected.

III. IMPACT OF BIT-FLIP COMBINATIONS ON THE DECODER

Increase in the size of the bit-flip combinations results in an increase in the decoding capability, by considering a larger set of errors. For bf = 8 bits, a set of 8 bit-flip errors are evaluated by the successive iterative decoder. Thus the error correction capability of the simple error-correction only RS decoder is increased up to an additional 8 symbols (if all the least reliable 8-bits belong to different symbols).

The error correction capability (E_{cc}) of the hard decision RS decoder is "t" i.e.,

$$E_{cc} = t \quad (4)$$

where,

$$t=(n-k)/2 \quad (5)$$

n is the codeword length and k is the dimension of the RS code, thus n-k is the number of parity symbols in the codeword. "t" is therefore half the number of parity symbols.

If the value of bf is increased from 8 to 16, this means the least reliable 16 bits will be considered in different combinations resulting in an increased decoding capability. If all the bits happen to be in different symbols then the decoding capability of RS codes can be extended up to bf-bits in total. With the successive iterative bit-flip decoding using a bit-flip combination of "bf" bits, the error correction capability (E_{ccbf}) can be given by the following equation,

$$Ecc_{bf} \leq t+bf, \quad 1 \leq bf \leq n*m \quad (6)$$

e.g., for bf = 8-bits an additional 8 symbol errors are correctable if all the 8 least reliable bits belong to different symbols. Similarly for bf = 16 bits, the additional error correctional capability is increased by 16 symbols (again if all the least reliable bits belong to different symbols). In the worst case the error correction capability is enhanced by one symbol. This happens when all the least reliable bits belong to the same symbol. However by increasing the bf, the chances of all the bits belonging to the same symbol decreases.

It is, however, infeasible to increase the value of bf up to n*m. Due to practical limitations, a smaller threshold value needs to be chosen.

It is to be noted that the increase in the error correction capability is beyond the error correction capability of the Chase 2 decoding algorithm (which is 2t). The reason is the presence of the CRC code in the data part, which will prevent the decoding error (miscorrection) and bring the received word from the Hamming sphere of another valid codeword back to its original position (or in the worst case at least identify it as a decoding error).

This, however, comes at the cost of extra computations. By increasing the bit-flip combinations by one bit, i.e., from bf to bf+1, we are effectively raising the size of error location set from 2^{bf} to $2^{(bf+1)}$. This means, the number of iterations of the successive iterative decoder is doubled i.e.,

$$2^{(bf+1)} = 2*2^{bf} \quad (7)$$

Increasing the value of bf, also results in an increased miscorrection detection and avoidance capability. This is because by increasing the value of bf, there are more error locations to be tried and corrected. If a decoding error occurs, it is detected with the help of CRC and tried in the next iterations for correction. Due to increased iterations, there remains a good probability that the errors in the received word will be corrected.

 ISBN: 978-1-61208-126-7

IV. SIMULATION RESULTS

Simulations are performed over an Additive White Gaussian Noise (AWGN) channel. A convolutional encoder of rate-½ is used with BPSK modulation. RS codes of different block lengths (n) over the corresponding Galois Field (GF (2^m)) are considered in the simulations. For each of the symbol size (m), a corresponding size of CRC code is chosen so as to occupy exactly one symbol, e.g., for m=8 the CRC-8 (CCITT) is used and for m=6, the CRC-6 (ITU) is used. Use of the CRC effectively reduces the code rate to (k-1)/n instead of k/n. To compensate for the reduced rate and give a fair comparison, the shortened RS codes are plotted for comparisons, e.g., RS (15, 11) with a 4-bit CRC is plotted against the same rate RS (12, 8) code. Symbol Error Rate (SER) vs. E_b/N_0 curves are shown in Figs. 3-5 to demonstrate the results.

Convolutional codes are good at converting the random channel errors to burst errors and RS codes have the excellent ability of correcting the burst errors. Burst errors may belong to a group of bits making one symbol or a group of symbols. Because of the excellent ability of the RS codes, to correct the symbol errors and erasures, SER (instead of Bit Error Rate) is plotted vs. the SNR, for better comparisons. It is to be noted that an improvement in SER suggests an improvement in the Bit Error Rate as well. In each of the plotted curves, the value of bit-flip combinations (bfc) is kept at 8, 12 and 16. It can be noticed from the curves that the SER drops with the increase in the bit-flip combinations.

A. RS(15, 11) code

Fig. 3 shows the simulation comparison for iteratively bit-flipped decoded Convolutional/ RS code for RS (15, 11). The effective code rate is 10/15 (i.e., (k-1)/n) because of the CRC, so it is compared with the same rate RS (12, 8) code. The simulation results show a coding gain of 0.25 dB at SER of 10^{-4} using a bf of size 8, a gain of 0.4 dB at same SER using a bf of size 12 and a coding gain of 0.5 dB using a bf of size 16.

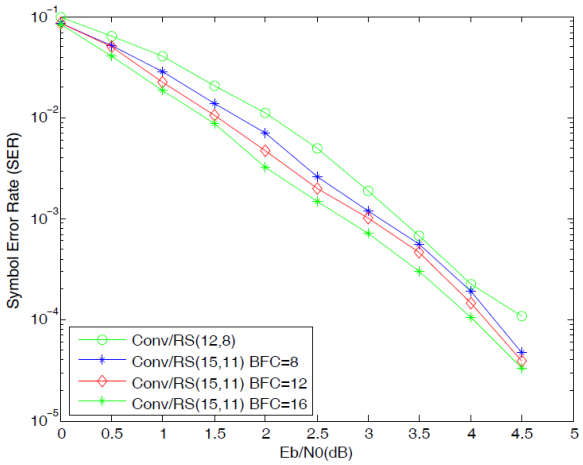


Figure 3. Convolutional/RS(15, 11) code with bit-flip combinations of 8, 12 and 16 bits compared to same rate RS(12, 8) code

B. RS(63, 55)

Fig. 4 shows the SER results for RS (63, 55) code compared with the same effective rate (i.e., 0.86) Reed Solomon code, i.e., RS (56, 48). A coding gain of 0.45 dB using a bfc of 8 bits is obtained at SER of 10^{-4}. A gain of 0.5 is obtained with a bf of size 12 and a gain of around 0.75 dB is obtained when 16-bit bit-flip combinations are considered.

C. RS(255, 223)

Fig. 5 shows an RS (255, 223) code with an 8-bit CRC, compared with the same effective rate RS(240, 208) code giving a coding gain of 0.2 dB using a bit-flip combination of 8 bits at SER of 10^{-3}. With 12-bit bit-flip combinations, the coding gain is 0.3dB and the coding gain increases to 0.4dB with a 16-bit bit-flip combinations.

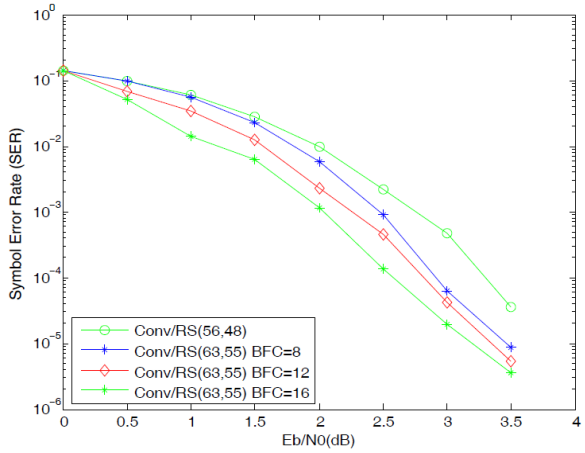


Figure 4. Convolutional/RS(63, 55) code with bit-flip combinations of 8, 12 and 16 bits compared to same rate RS(56, 48) code

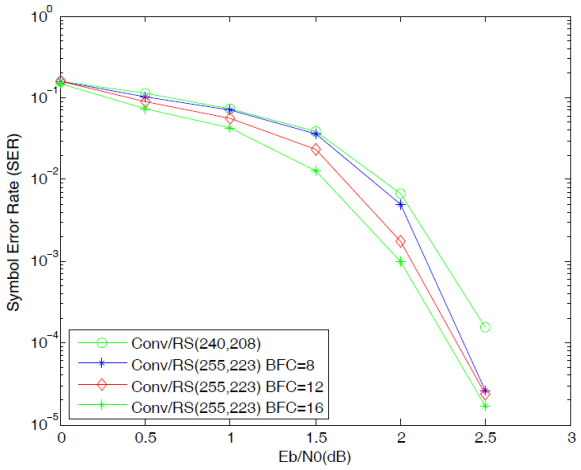


Figure 5. Convolutional/RS(255, 223) code with bit-flip combinations of 8, 12 and 16 bits compared to same rate RS(240, 208) code

In all of the simulations results, the curves for the different codes tend to meet with the curve for the hard decision Reed Solomon codes at the higher values of E_b/N_0. The reason for this is that at very high signal to noise ratio, the concatenated Convolutional/RS code is able to correct the errors in the received words with a high probability, thus giving almost the same SER as that of the bit-flipped scheme presented in this work. However the improvement in the SER is obvious for the smaller values of E_b/N_0 which is of more significance.

V. CONCLUSION AND FUTURE WORK

A scheme for the soft decoding of Reed Solomon codes is presented in this paper. The proposed scheme has the ability to correct errors beyond the decoding capability of Reed Solomon codes. The effect of the bit-flip combinations on the decoding scheme is analyzed. It is shown that by increasing the value of the threshold for the number of iterations to perform in the successive decoding of Reed Solomon codes, the error correcting capability increases and the probability of decoding error decreases. However, a very high threshold becomes practically infeasible.

The work presented in this paper is based on Error only Reed Solomon codes. In future it is planned to extend the work to Error and Erasure correcting Reed Solomon codes using symbol reliabilities in addition to the bit reliabilities. The error correction capability of the error and erasure RS codes is given by,

$$2*N_{err} + N_{era} <= 2t \quad (8)$$

where N_{err} is the number of errors and N_{era} is the number of erasures in the erroneous word. Using the symbol reliabilities, erasures can be introduced at known locations in the RS code and the decoding capability can be further enhanced by correcting up to 2t symbols. By being able to correct up to 2t erasures as well as the errors using the bit-flipping method discussed in this work, there is a likelihood of increasing the error correction capability beyond 2t.

REFERENCES

[1] O. U. Rehman and N. Zivic, Soft Input Decoding of Reed Solomon Codes with Miscorrection Detection and Avoidance, 4th International Conference on Signal Processing and Communication Systems, Gold Coast, Australia, pp. 1-5, Dec 13-15, 2010.

[2] G. David Forney, Concatenated Codes, MIT Press, Cambridge, MA, 1966.

[3] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, Optimal decoding of linear codes for minimizing symbol, IEEE Trans. Inform. Theory, vol. IT-20, pp. 284-287, Mar. 1974.

[4] R. J. McEliece and L. Swanson, On the decoder error probability for Reed-Solomon codes, IEEE Trans. Inform. Theory, vol. IT-32, pp. 701-703, 1986.

[5] D. Chase, A Class of Algorithms for Decoding Block Codes with Channel Measurement Information, IEEE Trans. Inform. Theory, IT-18, pp. 170-182, January 1972.

[6] G. David Forney, Generalized Minimum Distance decoding. IEEE Transactions on Information Theory, 12:125–131, 1966.

[7] D. J. C. MacKay and R. M. Neal, Good codes based on very sparse matrices, in Proc 5th IMA Conference Cryptography and Coding, (Lecture Notes in Computer Science), vol. 1025, pp. 100–111, 1995.

[8] M. Sudan, Decoding of Reed-Solomon beyond the error-correction bound, Journal of Complexity, vol. 13, pp. 180-193, Sep. 1997.

[9] R. Koetter and A. Vardy, Algebraic soft-decision decoding of Reed-Solomon codes, IEEE Trans. Information Theory, vol. 49, pp. 2809-2825, Nov. 2003.

[10] J. Jiang, Advanced Channel Coding Techniques using Bit Level Soft Information, PhD Thesis, Texas A & M University, Aug 2007.

 ISBN: 978-1-61208-126-7

Vulnerability of MRD-Code-Based Universal Secure Error-Correcting Network Codes under Time-Varying Jamming Links

Jun Kurihara*†
*KDDI R&D Laboratories, Inc.
2–1–15 Ohara, Fujimino,
Saitama, 356–8502 Japan.
Email: kurihara@kddilabs.jp

Tomohiko Uyematsu†
†Tokyo Institute of Technology
2–12–1 Ookayama, Meguro,
Tokyo, 152–8550 Japan.
Email: uyematsu@ieee.org

***Abstract*—In order to provide reliable and secure communication against eavesdroppers and jammers over networks, Universal Secure Error-Correcting Network Codes (USECNC) based on Maximum-Rank-Distance (MRD) codes have been introduced. This code can be applied to any underlying network codes. However, Shioji et al. introduced a reasonable network model against the code. In their model, an attacker eavesdrops information symbols from some links, where the set of eavesdropping links is re-selected during one packet transmission. The MRD-code-based USECNC cannot guarantee the security against eavesdroppers under this model. Inspired by Shioji et al.'s result, this paper considers the model such that the set of links that jamming (error) symbols are injected into is re-selected for each time slot. We show that the MRD-code-based USECNC cannot guarantee the error-correcting capability under the model of time-varying jamming links, even if the number of jamming links is limited to only one. Furthermore, by introducing a restriction on the field of local coding vectors in the network coding, we propose a simple solution to the problem of time-varying jamming links for MRD-code-based USECNC.**

***Keywords*-Network Coding; Secure Network Coding; Network Error-Correction; Jamming**

I. Introduction

Network coding [1] has been attracting much attentions since it can achieve better performance than ordinary networks with routing methods in terms of throughput, energy consumption, etc [2][3].

In reality, network coding may suffer from two kinds of adversaries: eavesdropping and jamming. Silva et al.'s presented Universal Secure Error-Correcting Network Codes (USECNC) [4][5] based on Maximum-Rank-Distance (MRD) codes [6][7] to provide secure and reliable communication against eavesdroppers and jammers over the network. Their code can be applied to any underlying network codes and hence it is called 'universal'. In the construction of Silva et al.'s USECNC, packets are split into m segments and transmitted to sink nodes through the network over m time slots. It has been assumed that packets are tapped by eavesdroppers and corrupted by jammers on several links that are fixed during one transmission of packets, i.e. m time slots.

Shioji et al. introduced the time-varying eavesdropping-link model in which attacker can re-select the set of eavesdropping links at each time slot of one packet transmission [8]. They showed that Silva et al.'s USECNC is insecure under this model, that is, information is leaked to eavesdroppers [8]. When the network coding is implemented on an overlay network of the Internet, the assumption of time-varying eavesdropping links is reasonable. Although a packet is not split from the perspective of the overlay network and they are transmitted through fixed 'logical' paths, a packet is split into multiple fragments and they are routed over different 'physical' paths at an intermediate router. Hence, from the point of view of the overlay network, the set of tapped logical links may be re-selected if physical links are eavesdropped. Such attackers against the network might be active, i.e. jammers who inject jamming packets into links to corrupt the network. Thus, this paper considers the model such that the attacker injects jamming (error) symbols into some links (called 'jamming links') in the network, and the set of jamming links is re-selected during one transmission of packets. We show that the MRD-code-based USECNC cannot guarantee the error-correcting capability under the model of time-varying jamming links, even if the number of jamming links is limited to only one. Furthermore, by introducing a restriction on the field of local coding vectors (LCV's) in the network coding, we propose a simple solution to the problem of time-varying jamming links for MRD-code-based USECNC.

The rest of this paper is organized as follows: Section II gives several definitions, the basic model of network coding and a brief review of MRD-code-based USECNC presented by Silva et al. In Section III, we introduce the time-varying jamming-link model, and show the vulnerability of MRD-code-based USECNC under this model. In Section IV, we propose a simple countermeasure against this model. Finally, we conclude this paper in Section V.

II. Preliminaries

In this section, we give definitions about the representation of finite field extensions, the basic model of network coding

 ISBN: 978-1-61208-126-7

and a brief review of MRD-code-based USECNC presented by Silva et al.

A. Field Extensions

Let $\mathbb{F}_q$ be a finite field containing q elements, and let $\mathbb{F}_{q^m}$ be an m-degree field extension of a base field $\mathbb{F}_q$. Then, $\mathbb{F}_{q^m}$ can be viewed as a vector space over $\mathbb{F}_q$. When the basis of the space is fixed, i.e., an irreducible polynomial generating the field extension is determined, an element of $\mathbb{F}_{q^m}$ can be represented by an m-dimensional vector over $\mathbb{F}_q$. We suppose that the vector representation of $x \in \mathbb{F}_{q^m}$ is written by $[x^{(1)}, x^{(2)}, \ldots, x^{(m)}] \in \mathbb{F}_q^m$.

B. Network Coding

Let $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ be a delay-free acyclic directed network, where $\mathcal{E}$ and $\mathcal{V}$ denote a set of links (edge, channel) and a set of nodes, respectively. Let $s \in \mathcal{V}$ and $\mathcal{R} \subset \mathcal{V}$ respectively denote a source node and a set of sink nodes, where $s \notin \mathcal{R}$. In this network model, we suppose that each link can carry an element of $\mathbb{F}_q$ per one time slot. The source node wishes to multicast the sequence $\vec{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}} \in \mathbb{F}_q^n$ to all sink nodes at rate n, where the rate is defined as the number of elements in $\mathbb{F}_q$ transmitted from s per one time slot. Suppose that

$$n \leq \min\{\mathsf{maxflow}(s, r) \ : \ r \in \mathcal{R}\}$$

holds, where $\mathsf{maxflow}(s, r)$ $(i, j \in \mathcal{V})$ denotes the max-flow from i to j. Then, there exists a network coding method in which s can multicast $\vec{x}$ to all nodes in $\mathcal{R}$ at rate equal to n [1].

We assume that linear network coding [9] is employed over $\mathcal{G}$, i.e., the type of data processing performed on the packets at each node is limited to linear combination. This implies that the data flow on any link over $\mathcal{G}$ can be represented as an $\mathbb{F}_q$-linear combination of the sequence $x_1, x_2, \ldots, x_n$. Thus, the information flow on link $e \in \mathcal{E}$ can be denoted as $y_e = \vec{b_e}^{\mathrm{T}} \cdot \vec{x}$ using a global coding vector (GCV), $\vec{b_e} = [b_1, b_2, \ldots, b_n]^{\mathrm{T}} \in \mathbb{F}_q^n$. When one has access to, say, l links $e_1, e_2, \ldots, e_l$, then the information obtained from these links is denoted as $M\vec{x} \in \mathbb{F}_q^l$, where $M = [\vec{b}_{e_1}, \vec{b}_{e_2}, \ldots, \vec{b}_{e_n}]^{\mathrm{T}}$. Constructing a network code is equivalent to determining the GCV of each link by setting the coefficients of the linear combination performed at each node.

C. The MRD-Code-Based USECNC

Here we introduce the fixed jamming-link model and the construction of Silva et al's USECNC employed over this model [4][5].

1) Fixed Jamming-Link Model: Suppose that one packet is composed of an element of $\mathbb{F}_{q^m}$ that is an m-degree field extension of $\mathbb{F}_q$, and is represented by an m-dimensional vector over $\mathbb{F}_q$. We define n packets transmitted from the source node s by $X = [X_1, X_2, \ldots, X_n]^{\mathrm{T}} \in \mathbb{F}_{q^m}^n (= \mathbb{F}_q^{n \times m})$. Then, the duration of one transmission of X is composed of m time slots. The source node s splits X and sends them over m time slots using a linear network code though $\mathcal{G}$, i.e., transmits $[X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)}]^{\mathrm{T}} \in \mathbb{F}_q^n$ at each time slot $i = 1, \ldots, m$.

Here we suppose that there exist $t (< n)$ jamming links in $\mathcal{E}$ and they inject error packets into $\mathcal{G}$. Silva et al. [4][5] assumed that these links are fixed during one transmission, i.e., m time slots. At a specific sink node, the received packets though $\mathcal{G}$ with injection of jamming (error) packets is represented by

$$Y = AX + DZ \in \mathbb{F}_q^{n \times m}, \tag{1}$$

where $A \in \mathbb{F}_q^{n \times n}$ is a transition matrix corresponding to GCV's. A linear network code employed over $\mathcal{G}$ is called feasible [10] if $\mathsf{rank}\ A = n$ for all sink nodes, otherwise it is rank-deficient. The rank deficiency is defined as $\rho = n - \mathsf{rank}\ A$. On the other hand, $Z \in \mathbb{F}_q^{t \times m}$ denotes t jamming (error) packets. $D \in \mathbb{F}_q^{n \times t}$ is a transition matrix corresponding to jamming links. D and Z are unknown random variables with unknown distributions. Then $\mathsf{rank}\ D \leq t$ holds since $t < n$ must hold.

Throughout this paper, we focus our attention only on injection of jamming packets, and omit eavesdropping on links (cf., Shioji et al.'s analysis in [8]).

2) Silva et al.'s Scheme: For the construction against the jamming(error)-packet injection, Silva et al.'s Universal Secure Error-Correcting Network Codes (USECNC) [4][5] is equivalent to $[n, k]$ Maximum Rank Distance (MRD) code $\mathcal{C} \subset \mathbb{F}_q^{n \times m}$ satisfying $m \geq n$ and $0 < k \leq n - 2t - \rho$, where ρ $(= n - \mathsf{rank}\ A)$ is the rank deficiency of A. Namely, the transmitted packet X in Eq.(1) is a codeword of $\mathcal{C}$. MRD code is a class of linear codes over $\mathbb{F}_{q^m}$, which is optimal in the rank-distance sense. For the $[n, k]$ MRD code $\mathcal{C}$, we have $d_R(C) \geq 2t + \rho + 1$, where $d_R(C)$ denotes the minimum rank-distance [6][7] between all pairs of distinct codewords of $\mathcal{C}$. We also have

$$\begin{aligned} d_R(AX, Y) &= \mathsf{rank}\ (Y - AX) \\ &= \mathsf{rank}\ DZ \\ &\leq \mathsf{rank}\ Z \\ &\leq t, \end{aligned} \tag{2}$$

where $d_R(P, Q)$ denotes the rank-distance between matrices $P, Q \in \mathbb{F}_q^{n \times m}$. Hence, the minimum rank-distance decoder of $\mathcal{C}$ can correct t jamming packets (and ρ rank deficiency of A) under the fixed jamming-link model $(n - k \geq 2t + \rho)$ [10].

III. Time-Varying Jamming Links over the Network

In this section, we introduce the model of time-varying jamming links, and show that MRD-code-based USECNC cannot guarantee the error-correcting capability under this model.

A. The Model

Suppose that the source node s multicast n packets, i.e., n vectors in $\mathbb{F}_q^m$, over time slots $i = 1, 2, \ldots, m$. We consider a model such that the jammer(s) can re-select the set of t jamming links at each time slot i.

Denote the t error packets injected into $\mathcal{G}$ by a $t \times m$ matrix

$$Z = [\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_m] \in \mathbb{F}_q^{t \times m},$$

where $\vec{z}_i \in \mathbb{F}_q^t$ $(i = 1, 2, \ldots, m)$ is a t-dimensional column vector chosen according to arbitrary distribution. At a specific sink node, let $D_i \in \mathbb{F}_q^{n \times t}$ $(i = 1, 2, \ldots, m)$ be a $n \times t$ matrix representing the linear combination of jamming-packet segments at time i. Jammer(s) specify the set of jamming-links arbitrarily, and hence we assume that D_i is chosen arbitrarily by jammer(s). The received matrix $Y \in \mathbb{F}_q^{n \times m}$ at the specific sink node is written as

$$Y = AX + V \in \mathbb{F}_q^{n \times m}, \tag{3}$$

where the matrix V denotes jamming packets conveyed over the time-varying jamming links, given by

$$V = [D_1\vec{z}_1, D_2\vec{z}_2, \ldots, D_m\vec{z}_m] \in \mathbb{F}_q^{n \times m}.$$

When $D_1 = D_2 = \cdots = D_m$, Eq.(3) is equivalent to Silva et al.'s fixed jamming-link model [4][5].

B. MRD-Code-Based USECNC under the Time-Varying Jamming Link Model

On the time-varying jamming-link model presented in the previous subsection, we first have the following lemma.

Lemma 1. *Suppose that the jammer(s) can arbitrarily select non-zero matrices $D_1, D_2, \ldots, D_m$ from $\mathbb{F}_q^{n \times t}$ and can generate non-zero vectors $\vec{z}_1, \vec{z}_2, \ldots, \vec{z}_m \in \mathbb{F}_q^t$. Then the maximum possible rank of V is n.*

Proof: It is only necessary to show an example having $\mathsf{rank}\ V = n$. Choose D_i for $i = 1, 2, \ldots, n$ $(n \leq m)$ as follows:

$$\begin{aligned} D_1 &= \left[\vec{u}_1, \vec{0}, \ldots, \vec{0}\right], \\ D_2 &= \left[\vec{u}_2, \vec{0}, \ldots, \vec{0}\right], \\ &\vdots \\ D_n &= \left[\vec{u}_n, \vec{0}, \ldots, \vec{0}\right], \end{aligned}$$

where $\vec{u}_j$ denotes the j-th column vector of an $n \times n$ identity matrix and $\vec{0}$ is a column zero vector. Let $\vec{z}_i$ be represented as $\vec{z}_i = [z_i^{(1)}, 0, \ldots, 0]^{\mathrm{T}} \in \mathbb{F}_q^t$. We assume that $z_i^{(1)}$ is chosen from $\mathbb{F}_q^* = \mathbb{F}_q \backslash \{0\}$ for $i = 1, \ldots, n$. Then, we denote V by

$$\begin{aligned} V &= [D_1\vec{z}_1, D_2\vec{z}_2, \ldots, D_n\vec{z}_n \mid D_{n+1}\vec{z}_{n+1}, \ldots, D_{m+1}\vec{z}_{m+1}] \\ &= [V_1 \mid V_2], \end{aligned}$$

where V_1 can be represented as

$$\begin{aligned} V_1 &= [D_1\vec{z}_1, D_2\vec{z}_2, \ldots, D_n\vec{z}_n] \\ &= \begin{bmatrix} z_1^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & z_n^{(1)} \end{bmatrix}. \end{aligned}$$

We thus have $\mathsf{rank}\ V_1 = n$ and hence $\mathsf{rank}\ V = n$ holds. This completes the lemma. ■

Denote $\mathsf{rank}\ V = v$. We then have $v \leq n$ from Lemma 1. Let X in Eq.(3) be a codeword of Silva et al.'s USECNC, that is, $[n, k]$ MRD code $\mathcal{C}$ with $m \geq n$ and $0 < k \leq n - 2t - \rho$. From Lemma 1, jammers may select D_i and $\vec{z}_i$ to satisfy $v > t$ even if $\mathsf{rank}\ D_i \leq t$ holds for all $i = 1, \ldots, m$. We then have

$$\begin{aligned} t < v &= \mathsf{rank}\ V \\ &= \mathsf{rank}\ (Y - AX) \\ &= d_R(Y, AX). \end{aligned}$$

Thus, the decoder of $\mathcal{C}$ cannot correct t jamming (error) packets under the time-varying jamming-link model.

Moreover, we give the following lemma that shows the number of jamming packets is independent of $\mathsf{rank}\ V$.

Lemma 2. *Suppose that jammer(s) can arbitrarily choose D_i's in Eq.(3) from $\mathbb{F}_q^{n \times t}$. Then, the maximum possible rank of V is n, which is independent of the value $t > 0$.*

Proof: Since the supposition $m \geq n$ is required in Silva et al.'s USECNC, the example presented in the proof of Lemma 1 always satisfy $\mathsf{rank}\ V = n$. Moreover, in the example in the proof of Lemma 1, we have assumed $\mathsf{rank}\ D_i = 1$ for $i = 1, 2, \ldots, n$ independently from the value $t > 0$. Thus the lemma is completed. ■

The above analysis proves the following result.

Theorem 1. *Suppose that there are t' jamming links in $\mathcal{E}$ under the time-varying jamming-link model. Let the transmitted packet over the network is a codeword of Silva et al.'s USECNC, i.e., $[n, k]$ MRD code $\mathcal{C}$ with $0 < k \leq n - 2t - \rho$ and $m \geq n$. Then, for any $t' > 0$, the decoder of $\mathcal{C}$ cannot correct t' jamming (error) packets injected from time-varying jamming links.*

This theorem implies that, under the time-varying jamming-link model, Silva et al.'s USECNC cannot guarantee the error-correction capability even if $t' = 1$.

C. An Example

Assume $q = 2$, $m = n = 3$ and $t = 1$. We do not consider the rank deficiency and the existence of eavesdroppers, and set $\rho = \mu = 0$. From these parameters, the USECNC is defined as a $[3, 1]$ MRD code $\mathcal{C} \subset \mathbb{F}_2^{3 \times 3}$. A codeword of $\mathcal{C}$ is defined as a 3×3 matrix $X = [X_1, X_2, X_3]^{\mathrm{T}} \in \mathcal{C}$, where $X_i \in \mathbb{F}_2^3$.

 ISBN: 978-1-61208-126-7

Under the time-varying jamming-link model, a specific sink node receives the following packet.

$$\begin{aligned} Y &= AX + V \\ &= A \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} + [D_1\vec{z}_1, D_2\vec{z}_2, D_3\vec{z}_3], \end{aligned}$$

where $A \in \mathbb{F}_2^{3\times 3}$, $D_i \in \mathbb{F}_2^{3\times 1}$ and $\vec{z}_i \in \mathbb{F}_2^1$ for $i = 1, 2, 3$. Here we assume $\mathsf{rank}\ A = 3$ and $\mathsf{rank}\ D_i \leq t = 1$ for $i = 1, 2, 3$. For example, we suppose A is a 3×3 identity matrix I and D_i's are

$$\begin{aligned} D_1 &= [1, 0, 0]^{\mathrm{T}}, \\ D_2 &= [0, 1, 0]^{\mathrm{T}}, \\ D_3 &= [0, 0, 1]^{\mathrm{T}}. \end{aligned}$$

We note that $\mathsf{rank}\ D_i = t = 1$ for each of $i = 1, 2, 3$ similar to the fixed jamming-link model. We also represent $X_i = [X_i^{(1)}, X_i^{(2)}, X_i^{(3)}] \in \mathbb{F}_2^3$ and $\vec{z}_i = [z_i] \in \mathbb{F}_2^1$ for $i = 1, 2, 3$. We then receive

$$Y = \begin{bmatrix} X_1^{(1)} & X_1^{(2)} & X_1^{(3)} \\ X_2^{(1)} & X_2^{(2)} & X_2^{(3)} \\ X_3^{(1)} & X_3^{(2)} & X_3^{(3)} \end{bmatrix} + \begin{bmatrix} z_1 & 0 & 0 \\ 0 & z_2 & 0 \\ 0 & 0 & z_3 \end{bmatrix}.$$

In this case, we have $\mathsf{rank}\ V = 3$ if all z_1, z_2, z_3 are nonzero and hence X cannot be uniquely reconstructed. This can be viewed as the worst case of Silva et al.'s USECNC under the fixed jamming-link model, that is, $3(> t)$ error packets are injected into the network.

IV. A Simple Way to Prevent the Jammers

The simplest way to prevent time-varying jamming links is to construct the network in which no packet is split into segments. Namely, packets are transmitted not as m-dimensional vectors using m time slots but as elements of an m-degree field extension itself at one time slot in Silva et al.'s USECNC. We thus consider how to transmit codewords of $[n, k]$ MRD code at one time slot over a linear network code and how to decode the received codeword by the minimum rank-distance decoder.

Let p be a prime and the number of elements in a base field $\mathbb{F}_p$. And let $p^m (\sim q)$ be the number of elements in an extended field $\mathbb{F}_{p^m}$. We redefine that each link can carry an element of $\mathbb{F}_{p^m}$ per one time slot. The source node s wishes to transmit $\vec{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}} \in \mathbb{F}_{p^m}^n$ to all sink nodes in $\mathcal{R}$ at one time slot using a linear network code. Then, the received packet at a specific sink node is represented as

$$\begin{aligned} \vec{y} &= [y_1, y_2, \ldots, y_n]^{\mathrm{T}} \\ &= A\vec{x} + D\vec{z} \in \mathbb{F}_{p^m}^n, \end{aligned} \tag{4}$$

where $A \in \mathbb{F}_{p^m}^{n\times n}$ is a transition matrix according to a linear network code over $\mathbb{F}_{p^m}$. $\vec{z} \in \mathbb{F}_{p^m}^t$ denotes t jamming packets and $D \in \mathbb{F}_{p^m}^{t\times t}$ is a transition matrix corresponding to jamming links.

Assume $\vec{x} \in \mathbb{F}_{p^m}^n (= \mathbb{F}_p^{n\times m})$ in Eq.(4) is a codeword of $[n, k]$ MRD code defined over $\mathbb{F}_{p^m}$, where $m \geq n$ and $0 < k \leq n - 2t + \rho$. Let $\mathcal{M}(\cdot)$ denote the matrix representation of a vector in $\mathbb{F}_{p^m}^n$ to measure the rank distance between vectors. For example, the matrix representation of $\vec{x}$ is

$$\mathcal{M}(\vec{x}) = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \end{bmatrix} \in \mathbb{F}_p^{n\times m},$$

where $[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}] \in \mathbb{F}_p^m$ is the vector representation of $x_i \in \mathbb{F}_{p^m}$ $(i = 1, \ldots, n)$. Then, according to Eq.(4), we have

$$\begin{aligned} d_R(\mathcal{M}(A\vec{x}), \mathcal{M}(\vec{y})) &= \mathsf{rank}\ (\mathcal{M}(\vec{y}) - \mathcal{M}(A\vec{x})) \\ &= \mathsf{rank}\ (\mathcal{M}(\vec{y} - A\vec{x})) \\ &= \mathsf{rank}\ \mathcal{M}(D\vec{z}), \end{aligned} \tag{5}$$

from the properties of the additive operation over a field extension $\mathbb{F}_{p^m}$. However, unlike Eq.(2), $\mathsf{rank}\ \mathcal{M}(D\vec{z}) \leq t$ does not always hold from the property of the multiplicative operation over $\mathbb{F}_{p^m}$. Hence, the minimum rank-distance decoder of the $[n, k]$ MRD code cannot be simply applied to Eq.(4). Hence, we add the following condition to the construction of the underlying network coding such that $\mathsf{rank}\ \mathcal{M}(D\vec{z}) \leq t$ always holds.

Condition 1. *For all nodes, elements of local coding vectors (LCV's) are chosen from the subfield (base field) $\mathbb{F}_p$ of the field extension $\mathbb{F}_{p^m}$.*

Note that elements of LCV's are coefficients of linear combination of incoming symbols at each node and they define elements of GCV's [2]. Thus, this condition also yields that the transition matrices A and D are defined over the subfield, i.e., $A \in \mathbb{F}_p^{n\times n} \subset \mathbb{F}_{p^m}^{n\times n}$ and $D \in \mathbb{F}_p^{t\times t} \subset \mathbb{F}_{p^m}^{t\times t}$.

For any $a \in \mathbb{F}_p$ and $b \in \mathbb{F}_{p^m}$, their vector representations are $[0, \ldots, 0, a] \in \mathbb{F}_p^m$ and $[b^{(1)}, b^{(2)}, \ldots, b^{(m)}] \in \mathbb{F}_p^m$, respectively. Then, the multiplication ab over $\mathbb{F}_{p^m}$ is given by

$$ab := [ab^{(1)}, ab^{(2)}, \ldots, ab^{(m)}],$$

from the fact of finite fields. That is, all operations can be executed over the base field $\mathbb{F}_p$. Thus, under Condition 1,

 ISBN: 978-1-61208-126-7

we have

$$\begin{aligned}
\mathcal{M}(D\vec{z}) &= \mathcal{M}\left(\begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,t} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ d_{t,1} & d_{t,2} & \cdots & d_{t,t} \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_t \end{bmatrix}\right) \\
&= \mathcal{M}\left(\begin{bmatrix} \sum_{w=1}^{t} d_{1,w} z_w \\ \sum_{w=1}^{t} d_{2,w} z_w \\ \vdots \\ \sum_{w=1}^{t} d_{t,w} z_w \end{bmatrix}\right) \\
&= \begin{bmatrix} \sum_{w=1}^{t} [d_{1,w} z_w^{(1)}, d_{1,w} z_w^{(2)}, \ldots, d_{1,w} z_w^{(m)}] \\ \sum_{w=1}^{t} [d_{2,w} z_w^{(1)}, d_{2,w} z_w^{(2)}, \ldots, d_{2,w} z_w^{(m)}] \\ \vdots \\ \sum_{w=1}^{t} [d_{t,w} z_w^{(1)}, d_{t,w} z_w^{(2)}, \ldots, d_{t,w} z_w^{(m)}] \end{bmatrix} \\
&= D \begin{bmatrix} z_1^{(1)} & \cdots & z_1^{(m)} \\ \vdots & \ddots & \vdots \\ z_t^{(1)} & \cdots & z_t^{(m)} \end{bmatrix} \\
&= D\mathcal{M}(\vec{z}),
\end{aligned}$$

where $d_{u,v} \in \mathbb{F}_p$ $(u, v = 1, \ldots, t)$ and $[z_w^{(1)}, z_w^{(2)}, \ldots, z_w^{(m)}] \in \mathbb{F}_p^m$ is the vector representation of $z_w \in \mathbb{F}_{p^m}$ $(w = 1, \ldots, t)$. Since $\mathsf{rank}\ \mathcal{M}(\vec{z}) \leq t$, we have

$$\mathsf{rank}\ \mathcal{M}(D\vec{z}) = \mathsf{rank}\ D\mathcal{M}(\vec{Z}) \leq t.$$

Combining Eq.(5) and this, the minimum rank-distance decoder of the $[n, k]$ MRD code can decode the received packets $\vec{y}$ as with the Silva et al.'s USECNC for the fixed jamming-link model.

However, this solution contains the following problems; a) The scheme is not universal since the condition of LCV's is introduced. b) Since the field for LCV's is changed from $\mathbb{F}_q$ to $\mathbb{F}_p$ with $p < q$, the probability of rank deficiency becomes high when random network coding [11] is employed, or the size of $\mathbb{F}_p$ might be insufficient for the deterministic construction of network coding depending on the network structure [12]. We believe the completely different technique from MRD codes must be required to realize universal secure error-correcting codes against the time-varying jamming links.

V. Conclusion

This paper considered the reasonable network model such that the set of links that jamming (error) symbols are injected into is re-selected for each time slot. We revealed that the MRD-code-based USECNC cannot guarantee the error-correcting capability under the model of time-varying jamming links, even if the number of jamming links is limited to only one. Further, by introducing a severe restriction to the field of LCV's, we presented a simple way to avoid jamming with the time-varying jamming links for MRD-code-based USECNC.

References

[1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[2] C. Fragouli and E. Soljanin, *Network Coding Fundamentals*. Now Publishers, 2007.

[3] C. Fragouli and E. Soljanin, *Network Coding Applications*. Now Publishers, 2007.

[4] D. Silva and F. R. Kschischang, "Universal secure error-correcting schemes for network coding." arXiv:1001.3387v1, Jan. 2010. Appeared in IEEE International Symposium on Information Theory (ISIT) 2010. Available at http://arxiv.org/abs/1001.3387v1.

[5] D. Silva and F. R. Kschischang, "Universal secure network coding via rank-metric codes." arXiv:0809.3546v2, Apr. 2010. Available at http://arxiv.org/abs/0809.3546v2.

[6] E. M. Gabidulin, "Theory of codes with maximum rank distance," *Problems of Information Transmission*, vol. 21, no. 1, pp. 1–12, 1985.

[7] R. M. Roth, "Maximum-rank array codes and their application to crisscross error correction," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 328–336, 1991.

[8] E. Shioji, R. Matsumoto, and T. Uyematsu, "Vulnerability of MRD-code-based universal secure network coding against stronger eavesdroppers," *IEICE Transactions on Fundamentals*, vol. E93-A, no. 11, pp. 2026–2033, 2010.

[9] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 371–381, 2003.

[10] D. Silva and F. R. Kschischang, "On metrics for error correction in network coding," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5479–5490, 2009.

[11] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4413–4430, 2006.

[12] S. Jaggi, P. Sanders, P. A. Chou, M. Effros, S. Egner, K. Jain, and L. M. G. M. Tolhuizen, "Polynomial time algorithms for multicast network code construction," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1973–1982, 2005.

 ISBN: 978-1-61208-126-7

An Enhanced Distance Measuring Scheme for DS-UWB Radar Systems

Youngpo Lee, Junhwan Kim, Dahae Chong, and Seokho Yoon†
School of Information and Communication Engineering, Sungkyunkwan University, Suwon, Korea
Email: {leeyp204, ourlife3, lvjs1019, and †syoon}@skku.edu
†Corresponding author

***Abstract*—This paper proposes a novel distance measuring scheme based on repeated PN sequence for ultra wide-band vehicle radar systems. The proposed scheme measures the distance between a vehicle and an obstacle based on repeated use of a single short pseudo random sequence. Simulation results show that the proposed scheme provides a shorter mean distance measuring time while keeping a similar measurement error performance to that of the conventional scheme.**

***Keywords*-PN sequence; DS-UWB; distance measuring; SRR; vehicle radar**

I. INTRODUCTION

Recently, the vehicle radar system has attracted much interest for improved road safety, since it enables drivers to safely control their vehicles by providing them with information on the distance between the vehicles and obstacles [1]-[4]. The vehicle radar system can be classified into two categories: long-range radar (LRR) and short-range radar (SRR) systems [3]. Specifically, we are concerned with the SRR system since it has lower cost and higher distance accuracy than those of the LRR system [3].

Due to its high resolution requirement, the SRR system generally employs the ultra wide-band (UWB) signal [1], [3], and its distance measuring accuracy becomes higher as the length of the pseudo-noise (PN) sequence used in the UWB increases [5]-[7]. However, a long PN sequence leads to a long distance measuring time, and consequently, is not appropriate for the vehicle radar system requiring a short distance measuring time to avoid a collision among high-speed vehicles.

To tackle this problem, recently, a distance measuring method using multiple PN sequences with different lengths instead of a single long PN sequence was proposed [7]. In this paper, on the other hand, a novel distance measuring scheme based on repeated use of a single short PN sequence is proposed. The simulation results demonstrate that the proposed scheme offers a shorter mean distance measuring time while maintaining a similar measurement error performance to that of [7].

The remainder of this paper is organized as follows. Section II describes the signal model and proposed distance measuring scheme. Simulation results are presented in Section III. Finally, Section IV concludes this paper.

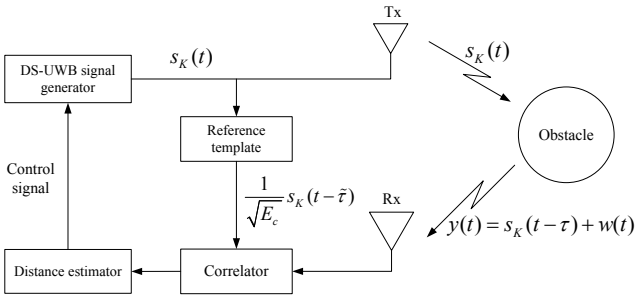


Figure 1. The system model of the proposed distance measuring scheme.

II. PROPOSED SCHEME

Fig. 1 shows the system model of the proposed distance measuring scheme. To measure the distance between a vehicle and an obstacle, first, the direct sequence (DS)-UWB signal generator transmits the following DS-UWB signal $s_K(t)$ to the obstacle:

$$s_K(t) = \sqrt{E_c} \sum_{j=0}^{KN-1} p_{\mathrm{mod}(j,N)} g(t - jT_c), \qquad (1)$$

where E_c and T_c are the chip energy and duration of a PN sequence with a length of N, respectively, K is the number of repeated use of the PN sequence, $p_j \in \{-1, +1\}$ is the jth chip, $\mathrm{mod}(\cdot,\cdot)$ is the modulo operation, and $g(t)$ is a UWB waveform with unit energy over $[0, T_c)$. Subsequently, the transmitted signal $s_K(t)$ is reflected from the obstacle and returns to the radar with a time delay τ, and thus, the received signal $y(t)$ can be expressed as

$$y(t) = s_K(t - \tau) + w(t), \qquad (2)$$

where $w(t)$ represents the additive white Gaussian noise (AWGN) process with double-sided power spectral density of $N_0/2$. To estimate the delay τ, next, a correlation between the received signal and a reference template $s_K(t-\tilde{\tau})/\sqrt{E_c}$, where $\tilde{\tau}$ is a trial value for τ, is performed as

$$\begin{aligned} R_K(\tau, \tilde{\tau}) &= \int_{t_1}^{t_1+KNT_c} \frac{1}{\sqrt{E_c}} y(t) s_K(t - \tilde{\tau}) dt \\ &= \int_{t_1}^{t_1+KNT_c} \frac{1}{\sqrt{E_c}} \{ s_K(t - \tilde{\tau}) s_K(t - \tau) \\ &\quad + s_K(t - \tilde{\tau}) w(t) \} dt \end{aligned} \qquad (3)$$

 ISBN: 978-1-61208-126-7

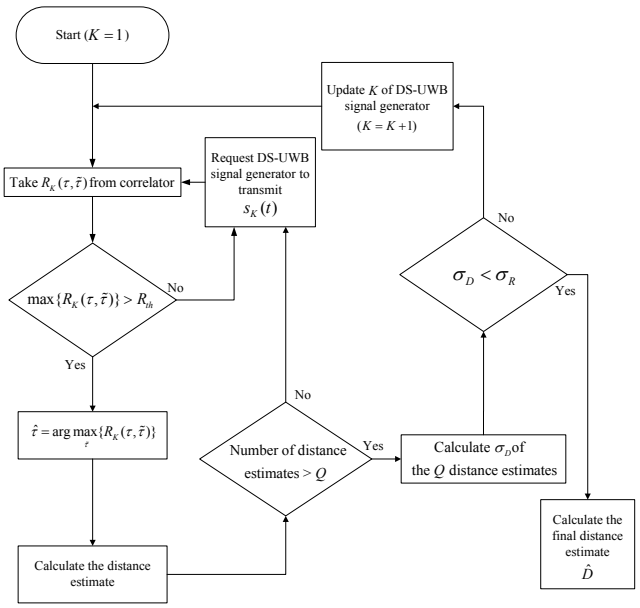


Figure 2. The algorithm of the distance estimator.

for $\tilde{\tau} = 0,\ T_c,\ 2T_c,\ \cdots,\ KNT_c$, where t_1 denotes the initial time of correlation. Finally, the distance estimator block estimates the distance between the vehicle and obstacle based on $R_K(\tau, \tilde{\tau})$, whose detailed process is shown in Fig. 2.

The maximum correlation value of $R_K(\tau, \tilde{\tau})$ is first selected and then compared with a threshold R_{th}. If the maximum value is larger than R_{th}, the delay estimate $\hat{\tau}$ is obtained as

$$\hat{\tau} = \arg\max_{\tilde{\tau}}\{R_K(\tau, \tilde{\tau})\}, \tag{4}$$

otherwise, the selection and comparison process resumes with newly transmitted $s_K(t)$. Once a delay estimate is obtained, the estimate of the distance D between the radar and obstacle is calculated as $\frac{c\hat{\tau}}{2}$, where c is the velocity of light (3×10^8 m/sec). To get a more reliable distance estimate, Q delay estimates and the corresponding Q distance estimates are obtained, and then, the standard deviation σ_D of the Q distance estimates is compared with a threshold σ_R specified by the system. If σ_D is less than σ_R, the final distance estimate $\hat{D}$ is obtained by averaging the Q distance estimates; otherwise, the overall distance estimation process is repeated with increasing K by 1. It should be noted that the proposed scheme is based on $R_K(\tau, \tilde{\tau})$, R_{th}, and σ_R independent from the speed of vehicles, and thus, valid for any speed of vehicles.

III. Simulation Results

In this section, the proposed scheme is compared with the conventional scheme [7] in terms of the mean distance measuring time and mean distance measuring error defined as the mean time that elapses prior to obtaining $\hat{D}$ satisfying $\sigma_D < \sigma_R$ and $E\{D - \hat{D}\}$, respectively. For simulations, we assume the following parameters: $N = 15$, $Q = 100$, $\sigma_R = 3$ m (10% of the maximum detectable range 30 m [2] of the SRR systems), and $T_c = 2$ ns. R_{th} is set with a false alarm probability of 0.01, and D is assumed to be distributed uniformly over the detectable range $[5, 30]$ of the SRR systems. For the conventional scheme, we use PN sequences with lengths of 15, 30, 60, 120, and 240 as in [7]

Fig. 3 shows the mean distance measuring time of the proposed and conventional schemes as a function of E_c/N_0. From the figure, the proposed scheme is observed to outperform the conventional one as expected. This is due to the fact that the proposed scheme uses a single short PN sequence repeatedly unlike the conventional scheme employing multiple PN sequences with different lengths.

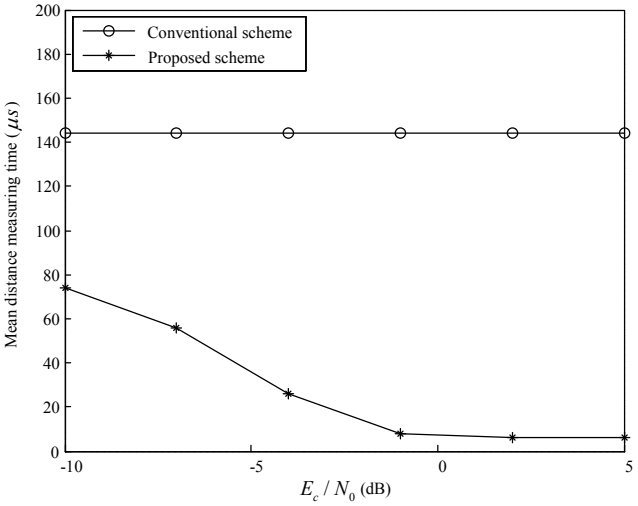


Figure 3. The mean distance measuring time of the proposed and conventional schemes.

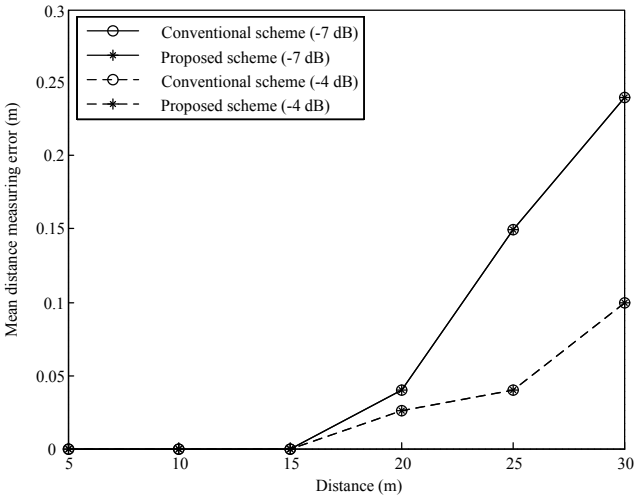


Figure 4. The mean distance measuring error of the proposed and conventional schemes at E_c/N_0 of -7 dB and -4 dB.

Fig. 4 shows the mean distance measuring error performance of the proposed and conventional schemes when E_c/N_0 is -7 dB and -4 dB. As we can see from the figure, the proposed scheme exhibits a similar measurement error performance to that of the conventional scheme in the detectable range of the SRR systems.

IV. Conclusion

In this paper, we have proposed a novel rapid distance measuring scheme for UWB vehicle radar systems, where a single short PN sequence is repeatedly used to yield a distance estimate, resulting in a shorter mean distance measuring time than that of the conventional scheme using multiple PN sequences with different lengths. The simulation results confirm that the proposed scheme outperforms the conventional scheme in terms of the mean distance measuring time while exhibiting a similar measurement error performance to that of the conventional scheme in the detectable range of the SRR systems.

Acknowledgment

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2010-0014610 with funding from the Ministry of Education, Science and Technology (MEST), Korea, and by the Information Technology Research Center program of the National IT Industry Promotion Agency under Grant NIPA-2010-C1090-1011-0005 with funding from the Ministry of Knowledge Economy, Korea.

References

[1] I. Gresham, A. Jenkins, R. Egri, C. Eswarappa, N. Kinayman, N. Jain, R. Anderson, F. Kolak, R. Wonhlert, S. P. Bawell, J. Bennelt, and J. P. Lanteri, "Ultra-wideband radar sensors for short-range vehicular applications," *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 9, pp. 2105-2122, Sep. 2004.

[2] K. M. Strohm, H.-L. Bloecher, R. Schneider, and J. Wenger, "Development of future short range radar technology," in *Proc. European Radar Conf. (EuRAD)*, pp. 165-168, Paris, France, Oct. 2005.

[3] J. Wenger, "Automotive radar - status and perspectives," in *Proc. Compound Semiconductor Integrated Circuit Symp. (CSIC)*, pp. 21-24, Palm Springs, CA, Nov. 2005.

[4] A. Komijani and A. Hajimiri, "A wideband 77-GHz, 17.5-dBm fully integrated power amplifier in silicon," *IEEE J. Solid-state Circuits*, vol. 41, no. 8, pp. 1749-1756, Aug. 2006.

[5] J. D. Taylor, *Introduction to Ultra-Wideband Radar Systems*, 1st ed. CRC Press, 2000.

[6] S. J. Xu, Y. Chen, and P. Zhang, "Integrated radar and communication based on DS-UWB," in *Proc. Int. Conf. Ultrawideband and Ultrashort Impulse Signals (UWBUSIS)*, pp. 142-144, Sevastopol, Ukraine, Sep. 2006.

[7] Y. Nakayama and R. Kohno, "Novel variable spreading sequence length system for improving the processing speed of DS-UWB radar," in *Proc. Int. Conf. Intel. Transport Sys. Telecommun. (ITST)*, pp. 357-361, Phuket, Thailand, Oct. 2008.

 ISBN: 978-1-61208-126-7

A Robust Periodogram-based IFO Estimation Scheme for OFDM-based Wireless Systems

Seung Goo Kang, Dahae Chong, Youngyoon Lee, and Seokho Yoon†
School of Information and Communication Engineering, Sungkyunkwan University, Suwon, Korea
Email: {general85, lvjs1019, news8876, and †syoon}@skku.edu
†Corresponding author

***Abstract*—In this paper, we propose a new periodogram-based integer frequency offset (IFO) estimation scheme robust to the fractional frequency offset (FFO) variation. We first observe the reason why the conventional IFO estimation scheme in [10] is sensitive to the variation of the FFO, and then, propose a new IFO estimation scheme using the modified maximum-likelihood (ML) metric. The numerical results demonstrate that the proposed scheme is more robust to the variation of the FFO and has better IFO estimation performance than the conventional scheme in [10].**

***Keywords*-estimation; frequency offset; OFDM; training symbol**

I. INTRODUCTION

Due to its immunity to multipath fading and high spectral efficiency, orthogonal frequency division multiplexing (OFDM) has been adopted as a modulation format in a wide variety of wireless systems such as digital video broadcasting-terrestrial (DVB-T), wireless local area network (WLAN), and worldwide interoperability for microwave access (WiMAX) [1]-[4]. However, the OFDM is very sensitive to the frequency offset caused by Doppler shift or oscillator instabilities, and thus, the frequency offset estimation is one of the most important technical issues in OFDM-based wireless systems [5], [6].

Various schemes [7]-[9] on the frequency offset estimation have been proposed so far. Schmidl and Cox (SC) proposed a frequency offset estimation scheme using a training symbol with two identical halves [7], whose estimation range is equal to the sub-carrier spacing. In [8], a new frequency offset estimation scheme that utilizes a training symbol with more than two identical parts was proposed, increasing the estimation range twice that of the SC scheme. However, the optimality for the estimation accuracy was not considered in the scheme in [8]. With the maximum-likelihood (ML) criterion, in [9], the optimal scheme for frequency offset estimation was derived using the same training symbol as in [8]. The scheme in [9] offers high estimation accuracy with the same estimation range as in the scheme in [8]. However, these schemes require a special training symbol structure, thus decreasing the transmission efficiency.

Recently, in [10], a periodogram-based frequency offset estimation scheme was proposed, which has the estimation range as large as the bandwidth of the OFDM symbol while maintaining the same performance as those of the schemes based on training symbols. However, its estimation performance for the integer part of the frequency offset (normalized to the sub-carrier spacing) rapidly changes according to the value of the fractional part of the frequency offset, eventually resulting in a significant variation in the overall frequency offset estimation performance.

Thus, in this paper, we propose a new integer frequency offset (IFO) estimation scheme robust to the fractional frequency offset (FFO) variation. We first investigate the influence of the FFO on the IFO estimation scheme in [10], and then, propose a modified ML IFO estimation scheme. The numerical results show that the proposed IFO estimation scheme is more robust to the variation of the FFO and has better performance than the IFO estimation scheme in [10].

II. SIGNAL MODEL

The nth transmitted complex-valued OFDM sample $x(n)$ is generated by using the inverse fast Fourier transform (IFFT), and thus, can be expressed as

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

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

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

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

where $s(n) = \sum_{k=0}^{L-1} h_k x(n-k)$ is the signal component with the kth channel filter tap coefficient h_k and the channel memory size L, ε_I and ε_F represent the IFO and FFO normalized to the sub-carrier spacing $1/T$, respectively, and $w(n)$ is the complex-valued additive white Gaussian

 ISBN: 978-1-61208-126-7

noise (AWGN) sample with mean zero and variance $\sigma_w^2 = \mathbf{E}\{|w(n)|^2\}$, where $\mathbf{E}\{\cdot\}$ and $|\cdot|$ denote the expectation and absolute value operators, respectively. In this paper, we assume that the channel is static during one OFDM signal period and timing synchronization is perfect.

III. Proposed Scheme

A. Influence of the FFO on the IFO estimation

In [10], the estimates $\hat{\varepsilon}_I$ and $\hat{\varepsilon}_F$ of the IFO and FFO are obtained as

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

and

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

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

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

where $f_k \in \{-\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}-1\}$ is the kth IFO candidate and $e(n) = \frac{x(n)^*}{\|x(n)\|^2}$ with the complex conjugate '$*$' and Euclidean norm $\|\cdot\|$ is the envelope equalized processing factor employed to remove the data modulation effect.

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

Fig. 2 shows the IFO metric $\{I(f) + I(f+1)\}$ normalized to $N^2\|h_0\|^2$ as a function of the frequency $f \in [-N/2, N/2)$ for $\varepsilon_F = 0.4$ and 0.8 when $\varepsilon_I = 1$, $N = 8$, and the noise is absent, where '$\circ$' represents the IFO metric value corresponding to each f_k and the shaded region represents the range of $\hat{\varepsilon}_I \in (\varepsilon-1, \varepsilon]$ for a correct estimation of the FFO. In this paper, the correct estimation probability of the IFO is defined as the probability that the maximum IFO metric corresponds to f_k within the shaded region. From the Fig. 2, we can clearly see that the correct estimation probability of the IFO would be very sensitive to the variation of the FFO, since the ratio of the IFO metric value corresponding to f_k within the shaded region to the largest one among the IFO metric values corresponding to f_ks' outside the shaded region rapidly changes according to the value of the FFO: specifically, the ratio when $\varepsilon_F = 0.4$ is larger than that when $\varepsilon_F = 0.8$, and thus, the correct estimation probability of the IFO would be higher when $\varepsilon_F = 0.4$ than when $\varepsilon_F = 0.8$.

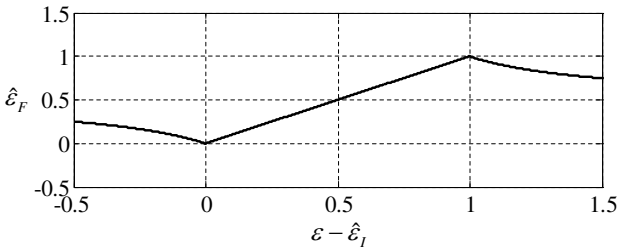
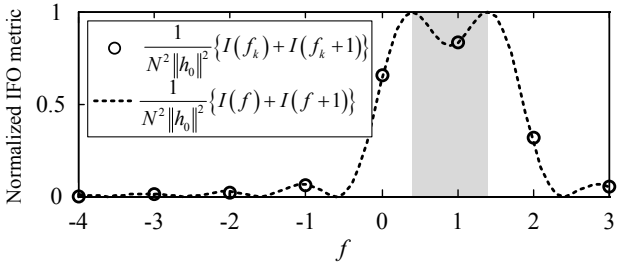
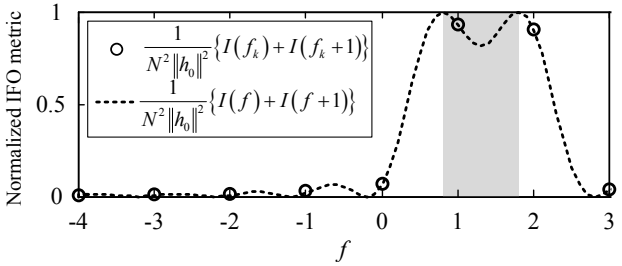


Figure 1. $\hat{\varepsilon}_F$ as a function of $\varepsilon - \hat{\varepsilon}_I$ in [10].



(a) When $\varepsilon_F = 0.4$



(b) When $\varepsilon_F = 0.8$

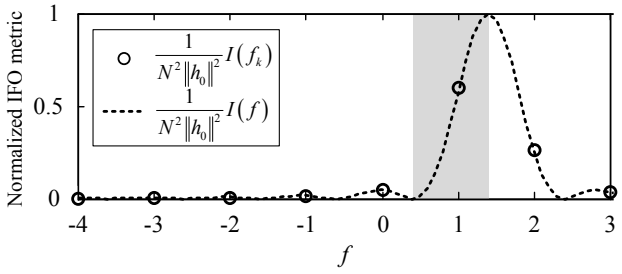
Figure 2. IFO metric $\{I(f) + I(f+1)\}$ normalized to $N^2\|h_0\|^2$ as a function of the frequency $f \in [-N/2, N/2)$ for $\varepsilon_F = 0.4$ and 0.8 when $\varepsilon_I = 1$, $N = 8$, and the noise is absent.

B. Proposed IFO Estimation Scheme

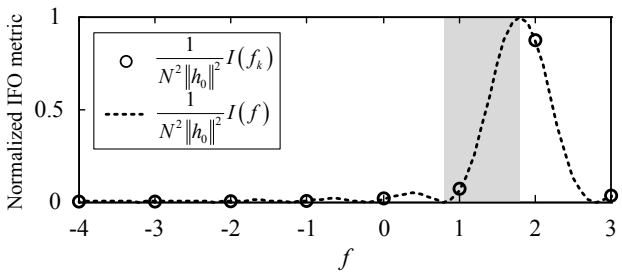
From the discussions in the previous section, we can see that the periodogram-based IFO estimation problem can be modeled as a detection problem of a single tone with the maximum energy. Then, the IFO estimate can be obtained as

$$\hat{\varepsilon}_I = \arg\max_{f_k} I(f_k), \tag{6}$$

which is in fact the ML solution for detecting a single tone with the maximum energy [11] and its normalized metric value is shown as a function of f in Fig. 3, where it is clearly observed that the correct estimation probability of the IFO would be still very sensitive to the FFO variation as that corresponding to (3), since (6) is an ML solution that does not take the FFO into account. Also, we can observe that the shaded region is always equal to the left slope of the normalized IFO metric graph in Fig. 3, and thus, through the some shift operation, the normalized IFO metric value in the shaded region can be increased regardless of the value

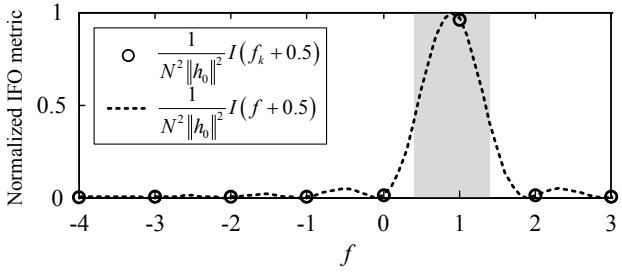


(a) When $\varepsilon_F = 0.4$

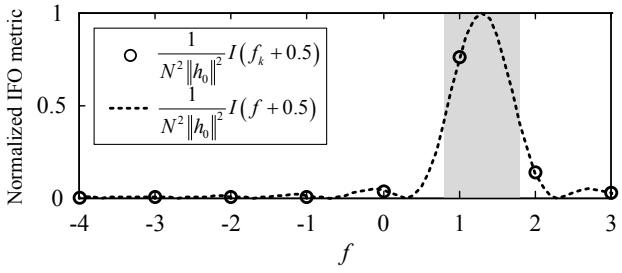


(b) When $\varepsilon_F = 0.8$

Figure 3. IFO metric $I(f)$ normalized to $N^2\|h_0\|^2$ as a function of the frequency $f \in [-N/2, N/2)$ for $\varepsilon_F = 0.4$ and 0.8 when $\varepsilon_I = 1$, $N = 8$, and the noise is absent.



(a) When $\varepsilon_F = 0.4$



(b) When $\varepsilon_F = 0.8$

Figure 4. IFO metric $I(f + 0.5)$ normalized to $N^2\|h_0\|^2$ as a function of the frequency $f \in [-N/2, N/2)$ for $\varepsilon_F = 0.4$ and 0.8 when $\varepsilon_I = 1$, $N = 8$, and the noise is absent.

of the FFO. Based on these observations, we propose the following modified ML IFO estimation scheme

$$\hat{\varepsilon}_I = \arg\max_{f_k} I(f_k + 0.5), \tag{7}$$

which is the shifted version of (6) to the left by 0.5 and its metric $I(f+0.5)$ always has the maximum value within the shaded region regardless of the value of the FFO as shown in Fig. 4 since the length of the shaded region is 1 and the FFO is distributed uniformly over $[0, 1)$. Thus, the correct estimation probability of the IFO in the proposed scheme is expected to be robust to the FFO variation.

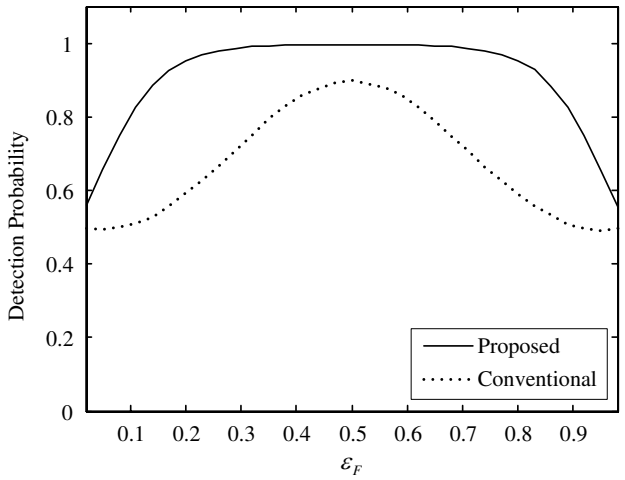


Figure 5. Correct estimation probabilities of the IFO as a function of the FFO for the proposed and conventional schemes in the AWGN channel model when SNR is 0 dB.

IV. Numerical Results

In this section, we compare the performance of the proposed IFO estimation scheme with the conventional scheme in [10]. In the simulation, we assume the following parameters: quadrature PSK (QPSK) modulated data sequence $\{X_k\}_{k=0}^{N-1}$, the FFT size of $N = 64$, a CP with a length of 8 samples, and the maximum Doppler shift of 125 Hz (corresponding to a mobile speed of 54 km/h and a carrier frequency of 2.5 GHz used for WiMAX). The signal to noise ratio (SNR) is defined as σ_s^2/σ_w^2 with $\sigma_s^2 \triangleq \mathbf{E}\{|s(n)|^2\}$. We consider AWGN and four-path Rayleigh fading channel models with path delays of 0, 2, 4, and 6 samples and exponential power delay profile of $\mathbf{E}\{A_l^2\} = \exp(-0.768l)$ (i.e., the power ratio of the first and last paths is set to be 10 dB).

Figs. 5 and 6 show the correct estimation probabilities of the IFO as a function of the FFO for the proposed and conventional schemes. As expected, the proposed scheme is more robust to the FFO variation than the conventional scheme. In addition, we can see that the proposed scheme significantly outperform the conventional scheme. This can be explained as follows. Since the proposed IFO metric is based on a single periodogram, it has only a single peak and can exploit it for detection. On the other hand, the conventional scheme uses the sum of two periodograms, and thus, has two peaks in the metric. The two peaks increase the metric value corresponding to the correct IFO candidate; however, they also increase the metric values

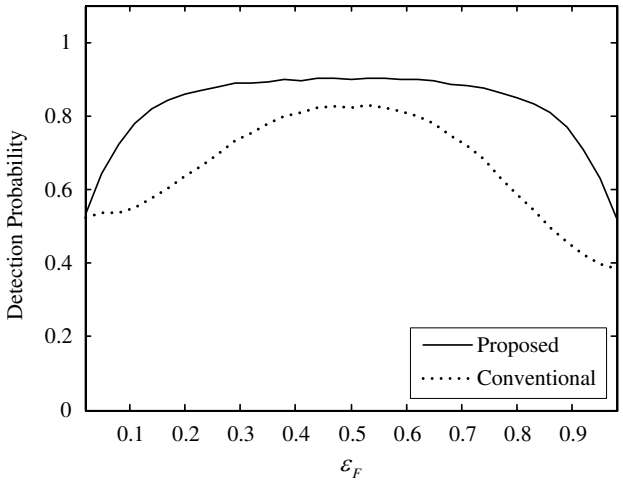


Figure 6. Correct estimation probabilities of the IFO as a function of the FFO for the proposed and conventional schemes in the Rayleigh fading channel model when SNR is 5 dB.

corresponding to the incorrect IFO candidates located on both sides of the correct IFO candidate, thus resulting in a higher incorrect estimation probability (i.e., a lower correct estimation probability) of the IFO.

V. Conclusion and Future Works

In this paper, we have proposed a new IFO estimation scheme robust to the FFO variation. We have first studied the influence of the FFO on the conventional IFO estimation scheme in [10], and then, proposed a modified ML IFO estimation scheme. By using efficiently shifted version of ML metric, the proposed scheme always has the maximum metric within the correct estimation range regardless of the value of the FFO. From the numerical results, it is confirmed that the proposed scheme is more robust to the variation of the FFO and has better IFO estimation performance than the conventional scheme in [10].

Acknowledgment

This research was supported by the National Research Foundation (NRF) of Korea under Grant 2010-0014610 with funding from the Ministry of Education, Science and Technology (MEST), Korea, and by the Information Technology Research Center program of the National IT Industry Promotion Agency under Grant NIPA-2010-C1090-1011-0005 with funding from the Ministry of Knowledge Economy, Korea.

References

[1] K.-T. Lee and J.-S. Seo, "Pilot-aided iterative frequency offset estimation for digital video broadcasting (DVB) systems," *IEEE Trans. Consumer Electron.*, vol. 53, no. 1, pp. 11-16, Feb. 2007.

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

[3] A. Filippi and S. Serbetli, "OFDM symbol synchronization using frequency domain pilots in time domain," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 3240-3248, June 2009.

[4] K. Cai, X. Li, J. Du, Y.-C. Wu, and F. Gao, "CFO estimation in OFDM systems under timing and channel length uncertainties with model averaging," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 970-974, Mar. 2010.

[5] T. Pollet, M. V. Bladel, and M. Moeneclaey, "BER sensitivity of OFDM to carrier frequency offset and Wiener phase noise," *IEEE Trans. Commun.*, vol. 43, no. 2, pp. 191-193, Feb. 1995.

[6] M. Morelli and M. Moretti, "Fine carrier and sampling frequency synchronization in OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1514-1524, Apr. 2010.

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

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

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

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

[11] A. Milewski, "Periodic sequences with optimal properties for channel estimation and fast start-up equalization," *IBM J. Res. Develop.*, vol. 27, no. 5, pp. 426-431, Sep. 1983.

 ISBN: 978-1-61208-126-7

Testing Triple Play Services Over Open Source IMS Solution for Various Radio Access Networks

Haris Luckin
BH Telecom d.d. Sarajevo
Sarajevo, Bosnia and Herzegovina
haris.luckin@bhtelecom.ba

Mirko Skrbic
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
mirko.skrbic@etf.unsa.ba

***Abstract* — Nowadays, with the development of next generation networks, new standards for voice and video services have been introduced, and VoIP (*Voice over IP*) has become a very popular protocol. It leads to the convergence of networks which provide voice and data transfer services with enhanced and improved applications for customers. As a key element of the next generation networks, IMS (*IP Multimedia Subsystem*) architecture aims to offer multimedia services via fixed-mobile convergent solutions. In this paper, a work model offering triple play services via a UMTS network and an IMS subsystem is realized. This model is realized through an Open Source software solution for an IMS subsystem, and the success of the triple play service provided by this system is analyzed. Testing of access networks and of different codecs using an Open IMS Core system and a UMTS access network is carried out. The most successful codec is determined and proposed for use while offering triple play services. Further, an optimization of the UMTS radio-access network of a middle size operator is done by planning the optimal number of NodeB's for this software solution and the dimensioning and optimization of the same. The conclusion is that, from the standpoint of offering triple play services to a certain number of users, no significant extensions of the radio network capacity are needed. The resulting system for offering triple play services is functional and optimized for exploitation in an active network with high reliability.**

Keywords – UMTS; IMS; Open IMS Core; EDGE; WiFi.

I. INTRODUCTION

Third generation mobile systems (3G) were introduced in accordance with the needs of customers for high speed data transfers, in order to enable a broad spectrum of Internet services. The UMTS (*Universal Mobile Telecommunications System*) presents one of the most frequently used standards for third generation mobile networks [7]. The combination of Internet and mobile networks has resulted in UMTS enabling triple play services, whose primary goal is the transition of all current and future services to IP: data, voice and video.

IMS (*IP Multimedia Subsystem*) represents a standard that defines core and service layer architecture of the fixed and the mobile telecommunication networks for a new generation [8]. The specification of its main elements, protocols and mechanisms is based on the cooperation of leading telecommunication standardization groups: 3GPP, ETSI/TISPAN and IETF. The main technological characteristic of IMS is its transparency towards different access networks. This is why IMS is the main initiator of the evolution of telecommunication networks, leading towards fixed-mobile convergence (FMC) solutions. Users can access the integrated user interface and are offered a unique broad spectrum of services using different devices and access technologies.

In [1], Magedanz, Vingarzan and Harjoc tested the Open IMS Core system [9] when a million users access the same system. However, an analysis or evaluation of the system while providing triple play services in a corresponding real environment of a medium size operator is not presented. Vingarzan and Weik presented in [2], the performance results of the Open IMS Core system in the case of transferring signalization traffic for voice transmission over various access networks. Testing of the video service and choosing appropriate codecs is not observed. In [3], Prokkola, Perala, Hanski and Piri presented the testing of HSDPA and WCDMA networks. The performances of the HSDPA network are not compared with other wireless radio access technologies. The main research objectives presented in this paper are:

- Realization of the IMS test platform based completely on open source technology.
- Realization and testing of triple play services over an IMS platform (voice call, video call and data transfer).
- Comparison of the performance of different access technologies (EDGE, WiFi, UMTS) through an IMS test platform by measuring appropriate parameters.
- Selection of the optimal codec in order to achieve the best quality for voice and video communication.
- Optimization of the UMTS radio access network in terms of predicting the number of users who will initially use the triple play services over UMTS radio access networks and the Open IMS Core system, and finally the optimization of the number of NodeB's needed to meet the users' needs.
- Considering the results published in [1], the design of a potential IMS network for a medium size operator, which would meet the needs of the initial number of users and whose network core would be an Open IMS Core system.

In the next section, the testing of triple play services over various radio access networks (EDGE/WiFi/UMTS) is described, with an analysis of the quality of the voice, video and data services. Thereafter, the optimization of mobile networks is described. The analysis is performed based on data from a medium sized telecom operator with a developed mobile network (GSM/EDGE/UMTS) and a fixed network (PSTN/ADSL), assuming that the number of users of the mobile network is 1.000.000, and 700.000 users for the fixed network. This is followed by the

 ISBN: 978-1-61208-126-7

dimensioning and optimization of the Open IMS Core system.

II. TESTING OF THE IMS SYSTEM

The parameters observed for the testing of every access network are: packet loss, delay, jitter, MOS (*Mean Opinion Score*) and R factor. The jitter diagram and the value of the MOS factor, when testing the voice and video service, will be shown for each access network. All of these measurements will be done for the GSM FR (GSM Full Rate) codec. The presumption is that a unloaded Open IMS Core system is used. The Open IMS Core is installed on the PC with a 1024/128 Kbit/s access link, and the conversation participants are stationary.

Finally, these parameters will be tested for a UMTS access network using an iLBC (*Internet Low Bitrate Codec*) codec, and it will be determined if and what kind of progress was achieved by introducing this codec.

In the case of data and instant messaging transfer, it can generally be concluded that there were no problems in the realization of these services through any of the used radio access technologies.

A. Testing through the EDGE radio access network

The test results of the triple play services for the EDGE radio access will be listed below, with an analysis of the measured parameters.

Figure 1 shows the block diagram of the triple play testing scenario through the EDGE radio access network and the Open IMS Core system.

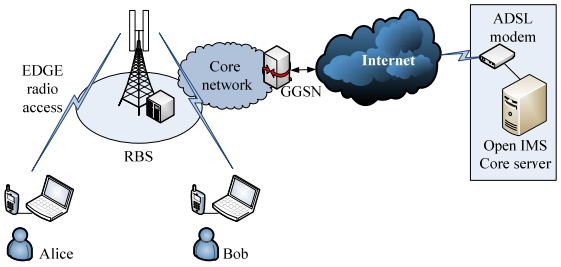


Figure 1. Testing of triple play services through the EDGE radio access network

The test results of the voice service through the EDGE radio access network and the Open IMS Core system can be interpreted as follows:

- The average value of the MOS parameter was 3.3, resulting in a very good quality of the voice communication.
- Variations in the packet delay were highly emphasized during the voice communication, as shown in Figure 2. This refers especially to the jitter increase in the interval between the 22nd and the 25th second of the voice session. Average jitter was 23.42 ms, and is marked by squares in the picture. According to [10], where the suggested maximum value for jitter in the VoIP network is 30 ms, it can be concluded that this network provides a good quality of the tested service.

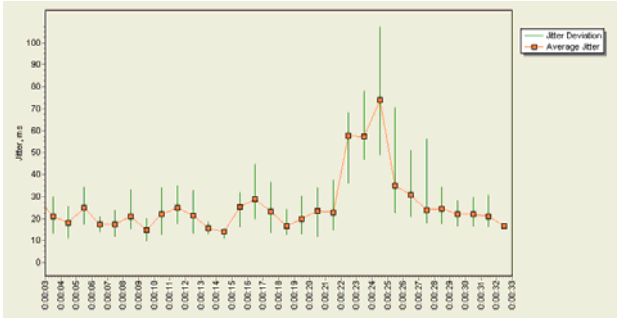


Figure 2. Measured jitter through the EDGE radio access network in the case of voice communication

While testing the voice and video service together, a large decrease in value of the MOS parameter during the session was noted. In the beginning of the communication, this parameter had the value of 3.7, then immediately decreased to 1 (which is basically an unusable call). It retains this value until the end of the call.

It can be seen from Figure 3 that the jitter had higher values during the first 9 seconds of the video session. Average jitter was 66 ms which is a very poor result for the tested type of service.

Figure 3. Measured jitter through the EDGE radio access network in the case of voice and video communication

From the measurements performed on the test system it can be summarized that video call services through an EDGE radio access network and an Open IMS Core system cannot be provided, since the introduction of a video call degrades the performances of the voice call to a level where it becomes impossible to use. This conclusion was to be expected given the practical transmission rate characteristic for EDGE technology.

B. Testing through the WiFi radio access network

The test results of the triple play services for a WiFi radio access network along with an analysis of the measured parameters will be presented below. Figure 4 shows the block diagram of the test scenario through a WiFi radio access network.

For the voice communication tested through the WiFi network, the MOS parameter had the value of 3.7, and was approximately stable during the entire period of the voice session, which is an excellent result for this type of codec.

 ISBN: 978-1-61208-126-7

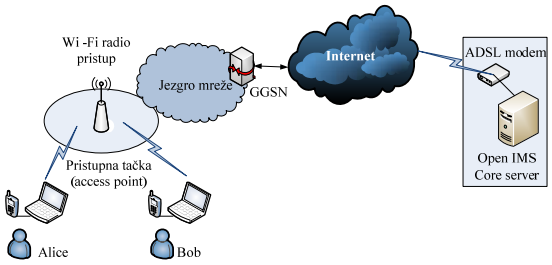


Figure 4. Testing of triple play services through a WiFi radio access network

Variations in packet delay were not highly expressed (as was the case for the EDGE access network) during the voice communication, as shown in Figure 5. Average jitter was 3.98 ms, which is a very good result for this type of service.

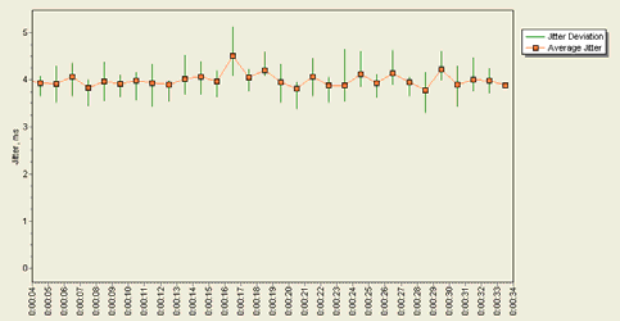


Figure 5. Measured jitter through the WiFi radio access network in the case of voice communication

While transferring voice and video communication through the WiFi network, the MOS parameter was 2.1, which means that the introduction of the video service significantly degraded the quality of the transmitted voice.

It can be concluded from Figure 6 that the jitter values during the call session were within acceptable limits. There was only one larger increase of these values in the 19th second of the call duration. Average jitter was 20.58 ms, which is an acceptable value for this type of service.

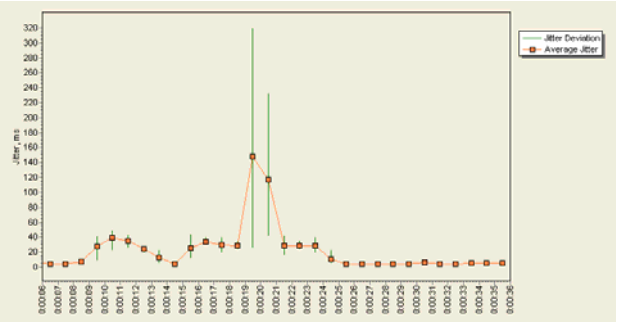


Figure 6. Measured jitter through the WiFi radio access network in the case of voice and video communication

It can be concluded that, when transmitting voice and video services over a WiFi access network and an Open IMS Core system, the transmission performance is significantly degraded when compared to the transmission of a voice service alone. The WiFi technology can be used for the transmission of video services, but a poorer quality can be expected for the enduser.

C. *Testing through a UMTS radio access network*

In this section, the test results of the triple play services for the UMTS radio access will be shown, with an analysis of the measured parameters.

Figure 7 shows a block diagram of the test scenario through the UMTS/HSxPA radio access network.

The test results of the voice service through the UMTS radio access network and Open IMS Core system can be interpreted as follows:

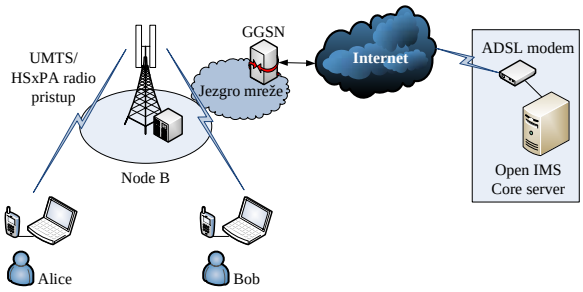


Figure 7. Testing of triple play services through UMTS radio access network

- MOS parameter of the voice service tested through the UMTS network was 3.7, which is, as we know, an excellent result. It is important to emphasize that this parameter was stable throughout the entire duration of the session.
- Variations in packet delay were not so emphasized during the voice communication, as is shown in Figure 8. Average jitter was 4.86 ms, which completely satisfies the requirements listed in [10].

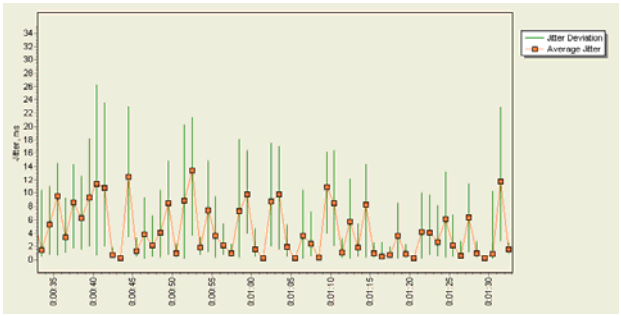


Figure 8. Measured jitter through the UMTS radio access network in the case of voice communication

The MOS parameter for the transfer of voice and video communication over the UMTS network was 1, which means that the introduction of the video signal notably degraded the quality of the voice transmitted; therefore, it cannot be used for those purposes.

From Figure 9 it can be concluded that jitter is emphasized at the beginning of the call, but later on during the communication, it decreases to a slightly lower

 ISBN: 978-1-61208-126-7

level. Average jitter was 48.29 ms, which is an insufficient value for this type of service.

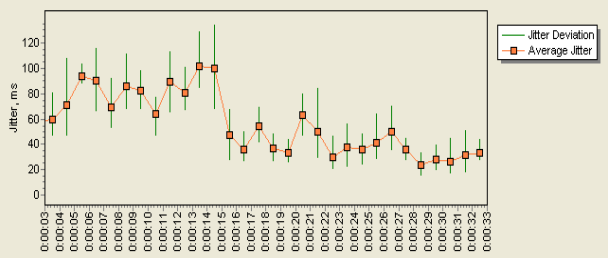


Figure 9. Measured jitter through the UMTS radio access network in the case of voice and video communication

It can be concluded that the transmission of a voice and video service through the UMTS access network and the Open IMS Core system significantly degraded the transmission performance, when compared to the transmission of voice service alone. Therefore, the UMTS technology cannot be used for the transmission of a voice and video service at the same time and under these conditions.

D. Testing through theUMTS radio access network – iLBC codec

In order to optimize and provide high quality triple play services, the analysis of parameters measured when using the iLBC codec in the network will be given below. This codec is primarily intended for the coding of speech in IP communications. The main advantage of this codec is an acceptable degradation of speech quality if it comes to packet loss or packet delay.

When testing the voice service using the iLBC codec through the UMTS access network and the Open IMS Core system, the results can be interpreted as follows:

- MOS parameter for the voice service tested through the UMTS network had an average value of 4.1, which is an excellent call quality. It also took stable values throughout the entire duration of the session.
- Variations in packet delay were not as emphasized during the voice communication, shown in Figure 10. Average jitter was 6.27 ms.

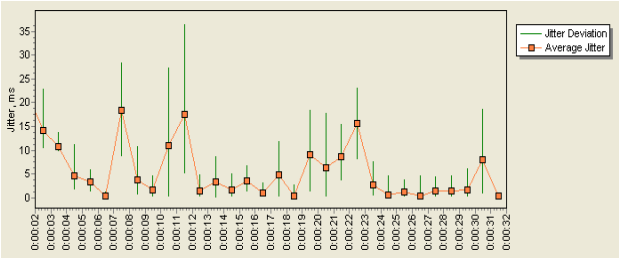


Figure 10. Measured jitter through the UMTS radio access network in the case of voice communication using iLBC codec

When testing the voice service together with the video signal through the UMTS access network and the Open IMS Core system, the results can be interpreted as follows:

- MOS parameter for voice and video communication through the UMTS access network had an average value of 3.9, and was stable throughout the entire session.
- Variations in packet delay were not as emphasized during the communication, except at the beginning of the call, as shown in Figure 11. Average jitter was 19.57 ms.

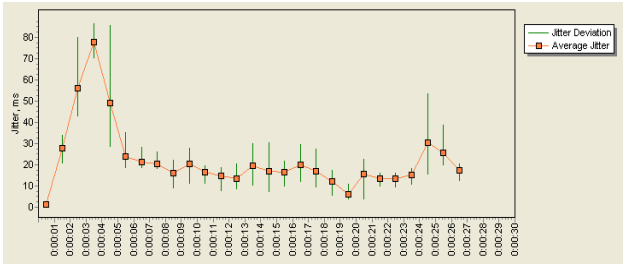


Figure 11. Measured jitter through the UMTS radio access network in the case of voice and video communication using iLBC codec

E. Conclusions

After obtaining the results of the transfer of the triple play service via the EDGE, UMTS and WiFi radio access networks, it can be concluded that the EDGE radio access network can be used for voice service, but the quality of the call is degraded completely when transferring the voice together with the video signal, and it cannot be used to offer these services to the end-users. As opposed to the EDGE network, the WiFi access network achieved an excellent performance for voice calls, however, for video calls there occurs a decrease in the quality of the call (MOS has the value 2), and it can be concluded that the WiFi technology, despite the significantly better performance, cannot be used for the transfer of triple play services over an Open IMS Core system, due to the relatively poor values of the MOS parameter. On the other hand, UMTS has also achieved an excellent performance for the voice call, while the video call has introduced a complete decrease in the quality of the service. It can also be concluded that UMTS is only to be used for voice and data services while it does not provide satisfactory parameters for a video service.

An improvement of the quality of voice and video communication through the Open IMS Core system and the UMTS radio access network using the iLBC codec is achieved in the final stage. When the iLBC codec is used to test the quality of triple play services of a voice communication, excellent performances of the quality of voice and video communication through the UMTS radio access network and the Open IMS Core system are achieved, which was to be expected taking into consideration the characteristics of this codec.

III. OPTIMIZATION OF THE RADIO ACCESS NETWORK

The fundamental factor for defining the number of users who will share network resources is the traffic

 ISBN: 978-1-61208-126-7

produced by each one of them. The traffic per user depends on the call frequency and average call duration.

One approach to the optimization of the UMTS radio access network for triple play users will be described below. Conditions, input data, presumptions and methodology for an optimal number of UMTS network NodeB's will be listed, taken from an average telecom operator.

Input data and presumptions are listed as follows:

- The operator has 1.000.000 mobile users.
- It is assumed that 20% of the mobile users use the UMTS service.
- It is assumed that 30% of the UMTS users use the triple play service.
- It is assumed that 10% of the triple play users access the system during busy hour.
- It is assumed that the RAB traffic class is used for the triple play service.
- Ideal propagation conditions are assumed.

RAB (Radio Access Bearer) represents a logical relation between the core network and user equipment. It is used to enable a connection for the UMTS service across the UTRAN (*UMTS Terrestrial Radio Access Network*) network. The RAB connections are realized as Radio Bearer connections between the RNC (*Radio Network Controller*) and the core network.

Triple play service users are considered to use the RAB interactive class of traffic and a certain service subscription package with a certain amount of traffic included. It is important to mention that (in order to define the busy hour) for packet traffic, it is assumed that 10% of the daily traffic is used during busy hour and the remaining 90% of data traffic is used during the rest of the day.

The traffic measured in kB/busy hour is calculated by dividing the amount of traffic from the subscription package into 30 days, taking into consideration ten percent of the obtained average daily traffic.

Considering the above calculation, and after the analysis of the average HSPA user, for the 1GB data traffic included in the monthly subscription, the traffic during busy hour is 28633 kbit.

The relation between uplink and downlink traffic is 10%, meaning that an average user realizes 318,19 kB of downlink traffic and 31,81 kB of uplink traffic during busy hour.

The configuration of every base station/cell will be designed to support HSDPA up to 14,4 Mbps and EUL up to 1,4 Mbps, using only one frequency bearer, because of the theoretical restriction of the code tree resources (HSDPA requires 15 out of the 16 available codes). This can only be realized with the assumption that there is no other (R99) traffic (otherwise the use of two bearers is necessary).

Since R99 traffic is still expected, it is optimal to include dynamical code allocation, i.e. a minimum of 10 codes per cell are allocated to the HSDPA, and the remaining five codes will be available when the R99 traffic is low.

If 10 codes per cell are allocated to the HSDPA, it can be concluded that the maximum traffic for that cell will be 7,2 Mb/s for downlink and 1,4 Mb/s for uplink traffic.

Considering the above assumption, that 10% of the triple play users will be accessing the network during busy hour, a total number of 600 active triple play users is obtained for busy hour.

For one cell it can be calculated:

- 7.2 Mbps/318,19 kB = 3 user/cell.

Therefore, it can be concluded that it is necessary to plan 200 cells/NodeB in the UTRAN network, so as to serve the total number of triple play users accessing the system during busy hour.

IV. DIMENSIONIG AND OPTIMISATION OF THE OPEN IMS CORE SYSTEM

In order to carry out the dimensioning and optimization of the Open IMS Core system, besides the number of UMTS users of the triple play service who will access the system during busy hour, it is necessary to define the number of fixed network users who will use the triple play service.

Input data and assumptions for this case are listed as follows:

- The operator has 700.000 users.
- It is assumed that 30% of the users use Internet service.
- It is assumed that 20% of the Internet users use triple play service.
- It is assumed that 10% of the triple play users access the network simultaneously during busy hour.

Using these assumptions a total number of 42.000 users who access the Open IMS Core system is obtained. Taking into account the assumption that 10% of the users use the triple play service at the same time, a total number of 4.200 users is obtained.

Together with the UMTS users, a number of 10.200 users accessing the Open IMS Core system at the same time is obtained.

Taking into account that the computer used for testing the Open IMS Core system has a poor performance, referring to the processor speed, the throughput of the network interface, memory, etc., this machine could not be used for an IMS core network of a real operator.

Such a workstation can, according to [1], simultaneously serve 10000 calls. This would satisfy the requirements listed above.

However, more users are expected to use the triple play service in the future, which will result in the need for proper dimensioning and optimization of the Open IMS Core system.

Scalability and redundancy are the next important parameters which should be observed. In order to achieve a scalable system, expandable in accordance with the future growth of the number of triple play users, and to improve the performances of the system for future users, two parallel Open IMS Core systems can be installed.

 ISBN: 978-1-61208-126-7

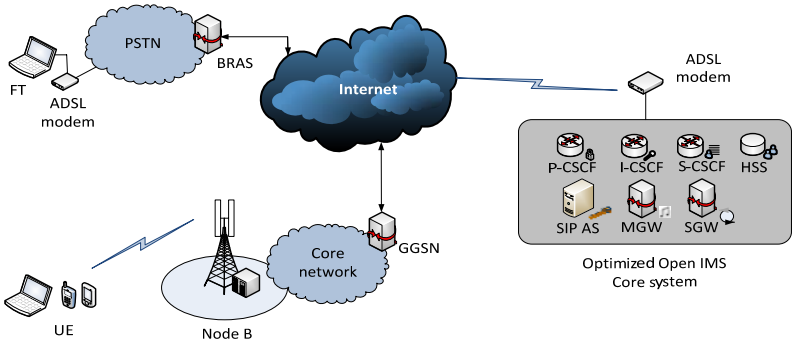


Figure 12. Optimized Open IMS Core system

This is the horizontal scalability principle, where a capacity-expandable platform is achieved by the multiplication of servers (CSCF's and FHoSS database).

From the viewpoint of system redundancy, it is possible to provide database redundancy, computer processor redundancy and geographical redundancy.

V. CHALLENGES FOR MOBILE OPERATORS – OTT (OVER-THE-TOP)

We are witnessing a large increase in mobile data traffic and a significant utilization of the available bandwidth of networks, currently happening worldwide. This growth receives great support from the diverse offerings of IP-enabled intelligent devices. To adjust and to earn additional revenue, mobile operators must optimize their networks. This optimization refers to the installation of new multimedia platforms in the core network (IP RAN aggregate node, IP core network). The new platforms would, in order to identify the used OTT services and to implement certain rules and repayments, enable the following functions:

- Service-aware charge.
- Access control.
- Policy control.
- Content filtering.
- Quality of service (QoS).
- Application detection and control.
- Optimization of data traffic.
- Security.

The Figure 13 presents these multimedia platforms.

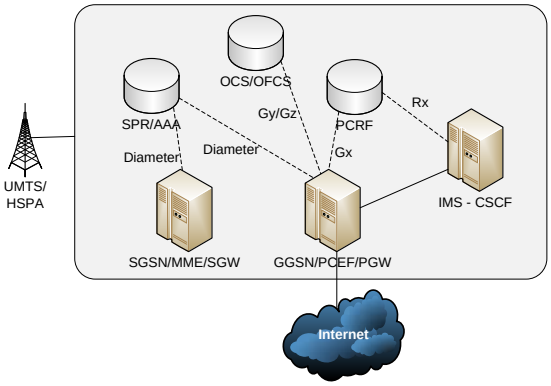


Figure 13. Multimedia platform

Through the implementation of the mentioned platforms, mobile operators get the opportunity to develop new models of revenue sharing with OTT providers. Extra revenue could be achieved by introducing new services.

VI. CONCLUSION

After obtaining the measuring results of the transfer of triple play services across EDGE, WiFi and UMTS access networks, it can be concluded that when introducing a video signal to a voice session over WiFi and UMTS access networks it shows a poor performance, which implies a significantly decreased quality of service. For the EDGE access network this service is completely degraded and is impossible to use.

For the voice transmission, very good performances are obtained for each of these three access technologies.

Further, it is experimented with using different codecs for the transmission over UMTS in order to improve the quality of service for this access network. The best results are obtained using the iLBC codec, which achieved great performances for triple play services.

From the viewpoint of data and instant messages transfer, it can generally be concluded that these sessions were successfully completed for each technology.

Further, the optimization of the number of UMTS radio access network's NodeB's is completed. This was done for a middle size operator, and the conclusion is that in order to offer triple play services to a certain number of users, no additional extensions of the radio access network are needed.

The increase in the number of users using advanced terminals and requests for the triple play service, results in an increase of the network load, which will result in an additional optimization of the network in the future.

Furthermore, a dimensioning of the IMS core for middle-sized fixed operators and mobile operators is done. It can be concluded that the Open IMS Core as a commercial solution would be a functional IMS system, adjusted to customer's requests, and with a high capacity, so user penetration could be satisfied for a long period of time.

Additional research could include testing of the WiFi network for different codecs, understanding jitter behaviour for characteristical measurements, and

 ISBN: 978-1-61208-126-7

measuring the same parameters over a longer measuring period.

REFERENCES

[1] T. Magedanz, D. Vingarzan, and B. Harjoc, "Generating Realistic NGN load with SIPNuke on Inexpensive Hardware – The One Million Demonstration" Fraunhofer FOKUS, Berlin, 2008.

[2] D. Vingarzan and P. Weik, "End-to-end Performance of the IP Multimedia Subsystem over Various Wireless Networks", Wireless Communications and Networking Conference, WCNC IEEE, 2006.

[3] J. Prokkola, P.H.J. Perala, M. Hanski, and E. Piri, "3G/HSPA Performance in Live Networks from the End User Perspective", IEEE International Conference on Communications, 2009.

[4] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality".

[5] ITU-T Recommendation G.107, "The E-model, a computational model for use in transmission planning".

[6] F. Yao and L. Zhang, "OpenIMS and Interoperability with Asterisk/Sip Express VOIP Enterprise Solutions", Agder University College, Faculty and Communication Technology, Grimstad, Norway, May, 2007.

[7] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "3rd Generation mobile system Release 1999 Specifications", 3GPP TS 21.101, last access date: 11.02.2011.

[8] 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, "IP Multimedia Subsystem (IMS), Stage 2, (Release 8)", 3GPP TS 23.228 V8.5.0 (2008-06), last access date: 11.02.2011.

[9] http://www.openimscore.org/, last access date: 11.02.2011.

[10] T. Szigeti and C. Hattingh, "End-to-End QoS Network Design", Cisco Press, 2004.

 ISBN: 978-1-61208-126-7

Novel Load Balancing Scheduling Algorithms for Wireless Sensor Networks

Endre László, Kálmán Tornai, Gergely Treplán
Faculty of Information Technology
Pazmany Peter Catholic University
Budapest, Hungary
lasen@digitus.itk.ppke.hu,
tornai.kalman@itk.ppke.hu, trege@itk.ppke.hu

János Levendovszky
Department of Telecommunications
Budapest University of Technology and Economics
Budapest, Hungary
levendov@hit.bme.hu

***Abstract*—In this paper, optimal scheduling mechanisms are developed for packet forwarding in wireless sensor networks, where clusterheads are gathering information. The objective is to monitor real-life processes for a given time interval and forward packets with minimum loss probabilities to the base station. In order to achieve this objective we develop an optimal scheduling algorithm, which determines the time slots in which packets must be sent by the nodes. The scheduling algorithm, on the one hand, guarantees that all the packets will be sent within a predefined time window and thus meeting delay constraints and, on the other hand, it provides uniform packet loss probabilities for all the nodes. The algorithm we propose is capable of providing optimal scheduling with given constraints and guarantees balanced load in polynomial time.**

***Index Terms*—load balancing, scheduling, wireless sensor networks (WSN), numerical optimization**

I. Introduction

Data gathering from a set of sensor nodes to a Base Station (BS) by using a cluster-based routing topology is commonly used in wireless sensor networks (WSNs) [1], [2]. In this kind of networks tiny sensor nodes communicate in short distances and collaboratively work to fulfill the application specific objectives of WSN. Many of the envisioned applications involve the collection of bursty data traffic generated by events, which are to be delivered to the BS as quickly and as reliably as possible in order to recognize emergency situation. In these applications packet delay and packet loss probability are of crucial importance [3], [4].

Resources (energy, time, bandwidth) are limited in WSN applications, hence cross-layer optimization are the key approach to minimize the utilization of resources if the Quality of Service (QoS) is prescribed.

Because of these strong limitations, scheduling becomes even more important in WSN to save on energy consumption. In this paper we develop a new scheduling algorithm for packet transfer in WSN, which guarantees reliable information transmission to the BS in terms minimizing the packet loss probabilities. Packet loss occurs when the amount of packets forwarded to the CHs exceed their capacity. Scheduling has been intensively researched in the telecommunication literature [5], [6], [7], however the main focus was on buffered architectures. The problem in our case will be transformed into a binary matrix optimization. So far, this has been solved by quadratic programming (see [8]). Now a novel solution is presented, which will provide the optimal scheduling matrix in polynomial time. Since clusterhead (CH) based routing is a commonly used solution in WSN (e.g., LEACH protocol [9] or other hierarchical solutions proposed in [10], [11]), we assume that each node can only send one packet to a selected CH at each time instant.

After the system and the application are briefly introduced, the network model and the mathematical model are detailed. The formulation and transformation of the original problem into a matrix optimization task is followed by the performance analysis and the conclusion of our work.

II. System description and application

The algorithm introduced in this paper is concerned with applications when the traffic generated by the sensor nodes are classified into traffic classes (e.g., motion detection, acoustic signals, video signals, ... etc.) The WSN architecture is single-tier clustered with heterogeneous sensors and performing centralized processing [12]. The CHs forward the received packets to the BS over a link with capacity V. As a result V presents a limit to the packets can be collected from nodes in a given time instant. In order to collect all the packets the CH assigns timeslots to the nodes for acquiring their packets. This timeslot assignment is referred as a scheduling. In this scheme, each traffic class has two parameters:

- the amount of packets to be transmitted to the CH;
- the time duration in which this given amount of packets must arrive at the CH.

The target platform is the Berkeley Mica2 mote [13], which is one of the most widely used WSN platforms. The platform has an 8 MHz processor, 4 kB of RAM, 128 kB of flash memory and a 433 MHz wireless radio transceiver. The transfer rate is 38.4 kbps and it is powered by two AA batteries.

TinyOS operating system [14], [15] is designed to be used with networked sensors and it supports the Mica2 platform. TinyOS handles task scheduling, radio communication, clock and timers, ADC and power management.

 ISBN: 978-1-61208-126-7

In the case of Mica2, a fully programmable Media Access Control (MAC) layer is available. A time division multiple access (TDMA) channelization is supported in the case of such simple and energy constrained nodes. Plenty of channel access method is proposed for WSN and there are two common solutions. One solution is the contention based scheme where nodes try to assign the channel randomly, independently from each other, such as ALOHA [16]. ALOHA is improved in many ways (such as B-MAC [17], Z-MAC [18], X-MAC [19]) to guarantee performance metrics required by different applications. Contention based protocols are very important channel access methods in the case of rare but bursty traffic, which is the typical case of event driven applications. On the other hand TDMA-based scheduled solutions could be very efficient if the users are known and fix, the data arrives regularly (e.g.: every one minutes the temperature value must be delivered to the BS) and topology is static. TDMA-based protocol can also save more energy, since each node can stay in sleep mode except for its own slot time. However TDMA-based protocols need synchronization protocols [20] affecting the guard time [21], time slot assignment [22] determining the end to end delay and scheduling. Scheduling protocols choose the order of the data transfer at the different links [23]. In this paper we propose Load Balancing Scheduling (LBS) algorithm, which minimizes the probability of cell loss.

In the following section the mathematical model will be introduced.

III. The model

The overall number of packets node j wants to send to the CH is denoted by X_j. All of the packets must be sent within a time interval $K_j : K_j > X_j$. The scheduling of packet transmissions by node j can be expressed by a binary vector $\mathbf{c}(j)$ of length K_j with weight X_j, where component $c_j(l) \in \{0, 1\}$ indicates whether packet is transmitted or not to the CH at time instance l.

Composing a transmission matrix $\mathbf{C}$ of these vectors, one may seek the optimal matrix ($\mathbf{C}_{\text{opt}}$). To find the optimal matrix, we present a novel approach based on a constructive iteration, solving the problem in polynomial time.

The data acquisition network can be divided into two parts:

- The sensors performing measurements in the different rooms.
- Relay nodes, which are forwarding the packets containing the measurements to the BS.

In Figure 1 the different application classes correspond to different rooms in a building. Each room has different X_j, i.e., one requests a given amount of data from every room. The data is collected by the sensor node from its sensors and stored for transmission. In the beginning of the communication X_j-s are transferred to the sensors nodes, where the calculation of K_j-s are performed based on the battery status and returned to the CH. The optimal scheduling matrix is then constructed by the CH. In each time-slot, V amount of data is sent to the CH corresponding to the l-th column of the scheduling matrix.

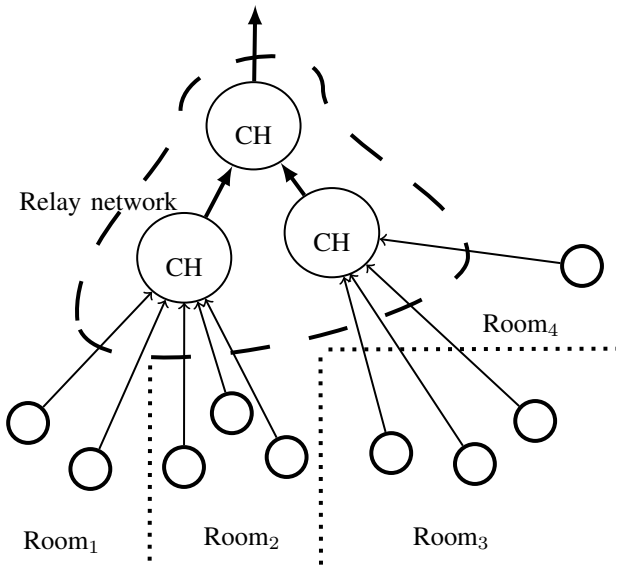


Fig. 1. WSN communication scenario with sensor nodes and CHs

This V amount of data is then sent to the relay nodes of the network in order to forward them to the BS.

The nodes performing the measurements are small size units with a variety of sensors included with strong energy and processing limitations. Energy is used either in stand-by mode or in active mode when they are engaged with radio transmission (packet forwarding). The available energy at node j is denoted by G_j, while the energy consumption in active node is g_{tx} and in stand-by is g_{sleep}. If the node sends X_j amount of packets and it switches off after K_j timeslot, then the overall energy consumption can be given as

$$G_j = (K_j - X_j) \cdot g_{\text{sleep}} + X_j \cdot g_{\text{tx}}. \quad (1)$$

As G_j is limited (due to the battery capacity) based on this expression one can determine what is the maximal K_j needed to transmit a given amount packets by the sensor node. As far as the relay nodes are concerned, they have more energy than the sensor nodes. Each room has a relay node, which acts as a CH. The CH can only receive V number of packets in a time frame due to its capacity. In the case of TDMA, V is the number of slots accommodated in a time frame. The task of scheduling is to assign time slots to the sensor nodes in such a way that the X_j number of packets are sent in K_j time and the probability of packet loss (in a given time slot more than V sensor nodes are sending packets to the CH) must be uniformly distributed.

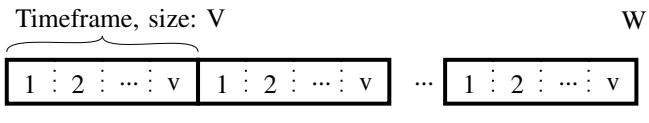


Fig. 2. Typical structure of TDMA timeframes

The number of timeframes available in a relay node is denoted by W. The number of available timeslots in each

timeframe is denoted by V. These slots are summarized on Figure 2. In a TDMA protocol each node has a dedicated timeslot in every timeframe wherein it may send its packet to its parent node.

A. Mathematical model

Let us assume that there are J number of nodes transmitting packets to a specific cluster head. The capacity of the cluster head in a time frame is denoted by V. This capacity means the amount of packet can be received in one time frame, the number of time slots. The amount of packets to be sent by node j is denoted by X_j, while the time constraint in which the transmission is to be finished is denoted by K_j. X_j is initially defined and using the equation (1) the K_j can easily be calculated.

The time is measured in discrete units thus $K_j, j = 1, \ldots, J$ are assumed to be integers. The scheduling of node j is represented by a binary vector $\mathbf{c}_j \in \{0,1\}^{K_j}$ where if $\mathbf{c}_j(l) = 1$ then a packet is sent to the cluster head at time instant l. The scheduling matrix $\mathbf{C}$ can be constructed from vectors $\mathbf{c}_j, j = 1, \ldots, J$, which form the row vectors of $\mathbf{C}$ and the number of columns is taken as $L = \arg\max_j K_j$. For example in the case of $J = 6$, $X_1 = 5$, $X_2 = 5$, $X_3 = 4$, $X_4 = 3$, $X_5 = 3$, $X_6 = 7$, $K_1 = 6$, $K_2 = 8$, $K_3 = 7$, $K_4 = 6$, $K_5 = 4$, and $K_6 = 8$ one specific scheduling matrix looks as follows:

$$\mathbf{C}_{\text{valid}} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

In this example the capacity of the CH was $V = 4$, and as it can be seen the number of ones are balanced. The following example for $\mathbf{C}$ matrix is not valid and unbalanced, and this scheduling overloads the CH.

$$\mathbf{C}_{\text{invalid}} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

The aggregated number of incoming packets in the cluster head at time instant l is given as $\sum_{j=1}^{J} c_j(l)$ being the weight of the l^{th} column vector in matrix $\mathbf{C}$. The cost of reception (which needs the power of CHs) is proportional to the number of received packets at time instant l expressed as $a_l := \sum_{j=1}^{J} c_j(l)$. We seek an optimal matrix $\mathbf{C}_{\text{opt}}$ where the aggregated number of incoming packets are balanced with respect to time. More precisely, we try to achieve that nodes schedule their packet transmission in such a way that each time instant the cluster heads receive more or less the same number of packets. This objective can be formulated as follows:

$$\mathbf{C}_{\text{opt}} : \min_{\mathbf{C}} \sum_{l=1}^{L} \sum_{k=1}^{L} \left(\sum_{j=1}^{J} C_{j,l} - \sum_{j=1}^{J} C_{j,k} \right)^2. \quad (2)$$

The constraint can be expressed as $\sum_{l=1}^{L} C_{j,l} = X_j, j = 1, \ldots, J$ and if the last nonzero component of row j is at location M_j, such that $M_j : \sum_{l=1}^{L} C_{j,l} = X_j, j = 1, \ldots, J$ then $M_j \leq K_j, j = 1, \ldots, J$. We seek the balanced solution by optimizing (2).

Furthermore we deal with the assumption that there is a loss in the transfer of packets to the CH because of the overload of the link. Namely, there exists no scheduling matrix for, which the aggregated load does not exceed the throughput V of the CH node, i.e., for every optimal scheduling matrix $\mathbf{C}_{\text{opt}}$ there exists at least one time instance when the aggregated load exceeds the throughput:

$$\forall \mathbf{C}_{\text{opt}} : \exists k, V < \sum_{j=1}^{J} X_j(k) \quad (3)$$

This assumption is important because it means that the available bandwidth is fully utilized.

B. The scheduling protocol

In order to optimize matrix $\mathbf{C}$ we need parameters $X_j, K_j, j = 1, \ldots, J$. In the first stage of the protocol we define the amount of packets we want to acquire data from the different sensor nodes (e.g. we want to get information about the status of the adequate room), which will specify parameters $X_j, j = 1, \ldots, J$. In the forthcoming discussion we define a communication protocol. Therefore, initially we assume that in each room $X_j, j = 1, \ldots, J$ are given.

1) The nodes are synchronized using the Reference Broadcast Synchonization (RBS [24]) protocol, in order to use the TDMA based package scheduling.
2) The values of X_j are sent to the sensor nodes in order to specify the amount of packets to obatin a certain amount of information about the status of room containing node j.
3) Each sensor node determines the amount of timeframes ($K_j, j = 1, \ldots, J$, one timeslot in each timeframe), wherein the X_j packets can be transmitted. In other word, by using equation (1) they calculate $K_j, j = 1, ..., J$. This information is sent back to the CH.
4) Having $X_j, K_j, j = 1, ..., J$ the CH will determine the optimal scheduling matrix $\mathbf{C}_{\text{opt}}$.
5) If at any given timeslot an overload occurs (the weight of a given column vector exceeds V) then some randomly selected packets are discarded subject to uniform distribution. In this way, the fair treatment is provided for all sensor nodes.
6) Each node will receive the corresponding row vector of the optimized scheduling matrix $\mathbf{C}_{\text{opt}}$.
7) In operation the nodes are transmitting packets to the CH only in those timeslot, which are indicated in the received row vector.

 ISBN: 978-1-61208-126-7

In the next section, the optimization of matrix $\mathbf{C}$ is discussed by using a constructive polynomial algorithm named LBS.

IV. Solution by polynomial algorithm: LBS algorithm

In this section a new algorithm is presented, which aims at balancing the number of "1"-s in every column vector of matrix $\mathbf{C}$. The forthcoming algorithm is motivated by a simple consideration that if the scheduling of the load can be distributed uniformly over a given time interval, then the packet loss is reduced to the minima. Any deviation from this uniformity may cause loss to a given time instant, while in other time slots under-utilization occurs.

Let the the matrix $\mathbf{C}$ initially be filled in with zeros. We modify the matrix as $C_{l,k} = 1$ if $k < K_l$ and $l = \arg\min_j K_j$. In this way, the algorithm will place L number if "1"-s in the matrix at the first L steps. One may see, that in each step that specific node is selected, which has the smallest remaining lifespan. After L steps the algorithm steps to the end of the matrix and the previous steps this construction procedure is repeated. Note that this algorithm will never put "1" in a place where the constraints forbid it ($X_j < K_j$). The algorithm terminates when all the prescribed $X_j, j = 1, ..., J$ number of packets are scheduled.

This method named LBS algorithm (Algorithm 1) can formally be described as follows:

Algorithm 1 LBS algorithm

Require: $\forall i = 1...J : K_i, X_i$
  $L \leftarrow \max_j K_j$
  $\mathbf{C} \leftarrow \mathbf{0}_{J \times L}$
  $S \leftarrow \sum_j X_j$
  while $S > 0$ **do** {Number of unscheduled packets}
    for $l \leftarrow L$ **to** 1 **do**
      Find $\arg\min_i K_i$ where $(l < K_i) \wedge (X_i > 0)$
      $S \leftarrow S - 1$
      $X_i \leftarrow X_i - 1$
      $\mathbf{C}_{i,l} \leftarrow 1$
    end for
  end while

Let us analyze the complexity of the newly proposed LBS algorithm. We put the number of $\sum_{j=1}^{J} X_j$ "ones" into the matrix and this can be upper bounded by JL. There is a minimization phase in each insertion, which be evaluated in time J. Thus, the overall complexity is as follows

$$J \cdot \sum_{j=1}^{J} X_j \sim \mathcal{O}\left(J \cdot (J \cdot L)\right) \sim \mathcal{O}\left(J^2 \cdot L\right). \qquad (4)$$

This complexity is between the cubic and quadratic time taking account the following assumption: $J < L$ therefore this algorithm runs in polynomial time.

The newly presented LBS algorithm gives optimal solution to (2) and runs in polynomial time. In the next section extensive simulations demonstrate the performance of the method.

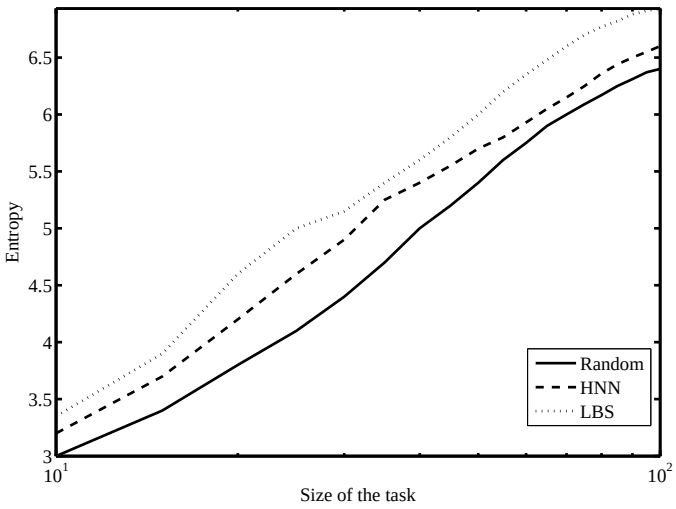


Fig. 3. Entropy of solutions achieved by different methods

V. Performance analysis

In this section, we compare the optimal solution obtained by exhaustive search to the one achieved by the LBS algorithm and to the one achieved by a random strategy. The simulations were carried out for $J = 20$ nodes and the corresponding $X_j, j = 1, ..., J$ and $K_j, j = 1, ..., J$ constraints have been chosen randomly in the range of $X_j \in [5, 100]$ and $K_j \in [5, 100]$. The results have been evaluated after selecting several constraints randomly, then running the simulations and taking the average error of the solutions achieved by the methods and by exhaustive search. The random method and the exhaustive search are the reference solutions in order to provide a better comparison of this new LBS method.

We took into consideration the properties of the platform specified in Section II in the simulations we performed.

The obtained results are depicted by the Figure 3. Analyzing the result on the figure, one may note that the solution provided by the LBS algorithm is possible as good as the results provided by the Exhaustive search. The Random strategy performs more worse than the LBS algorithm or the Exhaustive search. The figure shows the entropy of the solutions. One may note the horizontal axis scaled logarithmically. (The exhaustive search is not visible as the results of the LBS algorithm fully overlaps it.)

The previous and following graphs contain the results of a Hopfield Neural Network (HNN), our previous solution to solve the specified problem. Details about the HNN may be found in [8] and [25]. The results of the HNN only slightly differs from the optimum.

The running time of the enumerated methods are compare to each other in Figure 4. The LBS algorithm, the HNN and the random strategy provide its solutions in polynomial time opposed to to the exponential complexity of the exhaustive search method. (The exhaustive search due to the exponential complexity does not appear on the previous graph.) The LBS

 ISBN: 978-1-61208-126-7

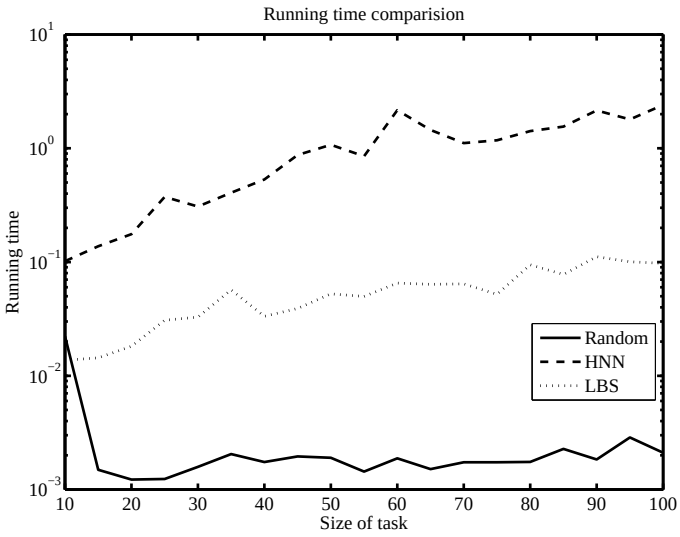


Fig. 4. Running time of solutions achieved by different methods

algorithm is faster than the HNN recursion, the reason are detailed:

- The random strategy generates a random matrix keeping the constraints $X_j, K_j : \forall 1 \dots J$. The complexity of this generation is
$$\mathcal{O}\left(J \cdot L\right)$$
- The theoretical complexity of the LBS algorithm is (as detailed previously):
$$\mathcal{O}\left(J^2 \cdot L\right)$$
- The time and steps of convergence for the discrete HNN using the strong Lyapunov criteria could be proven as follows:
$$\mathcal{O}\left(J^2 \cdot L^2\right)$$
- The exhaustive search uses huge search space and elects the optimal scheduling matrix from all possible $\mathbf{C}$ matrix. Therefore its complexity:
$$\mathcal{O}\left(2^{J \cdot L}\right)$$

Therefore we expect the results depicted by the Figure 4. (The running time is measured in seconds and the vertical axis is scaled logarithmically.) We repeated out tests several times but not enough times to provide monotone variation as one may expect. Certainly these results can be extrapolated.

In order to measure how balanced a solution is, we introduce the entropy of the weight distribution of the columns in matrix $\mathbf{C}$ as follows

$$H\left(\mathbf{p}\right) = \sum_{k=1}^{L} -p_k \ln p_k, \tag{5}$$

where

$$p_k = \frac{\sum_{j=1}^{J} C_{jk}}{\sum_{k=1}^{L}\sum_{j=1}^{J} C_{jk}}.$$

This entropy value provides a heuristic metric to describe the measure of equalization of the packet loads in the different solutions.

Table I demonstrates that the solution provided by LBS algorithm has the highest weight entropy (i.e., the most uniform weight distribution of the columns), hence it fulfills the constraint related to weight balancing.

Method	Entropy
Random Strategy	5.45
HNN	5.62
LBS algorithm	5.89
Exhaustive Search	5.89

TABLE I
ENTROPY OF WEIGHT DISTRIBUTION

Figure 5 shows the advantages of the LBS algorithm, since the number of packets hitting the CH at in time frames are equalized.

One may also see that if a capacity V is given (in this example let $V = 9$) then the scheduling methods provided either by the Random Strategy or by the HNN will suffer from packet loss, as the CH cannot receive all the packets in the given time instances. This effect is very sharp if the timeframes are less than 25. In the case of LBS algorithm, or Exhaustive Search there is no packet loss even in these cases due to more uniform distribution of the load.

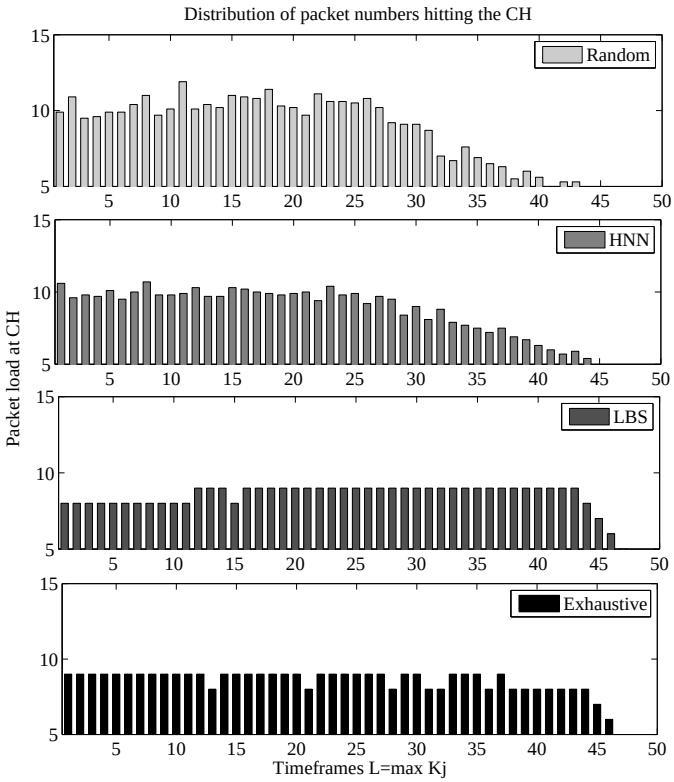


Fig. 5. Received packets in timeframes

The random strategy follows from a simple algorithm of constructing a scheduling matrix. In each row of the matrix the X_j number of "1"-s are randomly placed within the K_j time limit. This method is used for comparison, it serves as a "dummy" model.

VI. Conclusion

In this paper, new methods were proposed to provide optimal scheduling for packet-transmissions in WSN. The objective of optimal scheduling was defined by the Balanced Cost objective function. It is noteworthy, that the newly introduced methods yields always a valid scheduling matrix, as opposed to the HNN or random scheduling, which may yield invalid matrices. The algorithm introduced above is capable of providing optimal scheduling with given constraints and guarantees balanced load. It runs in polynomial time and yield valid solution therefore it is possible to use in real-life WSN data acquisition applications.

References

[1] A. Rogers, D. D. Corkill, and N. R. Jennings, "Agent technologies for sensor networks," *IEEE Intelligent Systems*, vol. 24, pp. 13–17, 2009.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.

[3] C.-Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1247–1256, 2003.

[4] P. Inverardi and L. Mostarda, "A distributed monitoring system for enhancing security and dependability at architectural level," pp. 210–236, 2007.

[5] O. Aumagie, E. Brunet, N. Furmento, and R. Namyst, "NewMadeleine: a Fast Communication Scheduling Engine for High Performance Networks," RR-1421, Tech. Rep., 2007.

[6] H. Choi, J. Wang, and E. A. Hughes, "Scheduling for information gathering on sensor network," *Wirel. Netw.*, vol. 15, no. 1, pp. 127–140, 2009.

[7] H. Li, P. Shenoy, and K. Ramamritham, "Scheduling messages with deadlines in multi-hop real-time sensor networks," in *RTAS '05: Proceedings of the 11th IEEE Real Time on Embedded Technology and Applications Symposium*. IEEE Computer Society, 2005, pp. 415–425.

[8] J. Levendovszky, E. Laszlo, K. Tornai, and G. Treplan, "Optimal pricing based resource management," in *International Conference on Operations Research Munich 2010*, September 2010, p. 169.

[9] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 8*. IEEE Computer Society, 2000.

[10] S.-H. Cha and M. Jo, "An energy-efficient clustering algorithm for large-scale wireless sensor networks," in *GPC'07: Proceedings of the 2nd international conference on Advances in grid and pervasive computing*. Springer-Verlag, 2007, pp. 436–446.

[11] K. Wu, C. Liu, Y. Xiao, and J. Liu, "Delay-constrained optimal data aggregation in hierarchical wireless sensor networks," *Mob. Netw. Appl.*, vol. 14, no. 5, pp. 571–589, 2009.

[12] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Comput. Netw.*, vol. 51, pp. 921–960, March 2007.

[13] CAPSIL. (2010) Mica2 technical parameters. [Online]. Available: http://www.capsil.org/capsilwiki/index.php/MICA2

[14] TOSWiki. (2010) Tinyos introduction. [Online]. Available: http://docs.tinyos.net/index.php/FAQ

[15] P. Levis. (2006) Tinyos programming. [Online]. Available: http://www.tinyos.net/tinyos-2.x/doc/pdf/tinyos-programming.pdf

[16] N. Abramson, "The aloha system - another alternative for computer communications," in *Proceedings of Fall Joint Computer Conference, AFIPS Conference*, 1970.

[17] J. Polastre, J. Hill, and D. Culler, "Versatile low power media access for wireless sensor networks," in *Proceedings of the 2nd international conference on Embedded networked sensor systems*, ser. SenSys '04. ACM, 2004, pp. 95–107.

[18] I. Rhee, A. Warrier, M. Aia, and J. Min, "Z-mac: a hybrid mac for wireless sensor networks," in *SenSys '05: Proceedings of the 3rd international conference on Embedded networked sensor systems*. New York, NY, USA: ACM Press, 2005, pp. 90–101.

[19] M. Buettner, G. V. Yee, E. Anderson, and R. Han, "X-mac: a short preamble mac protocol for duty-cycled wireless sensor networks," in *Proceedings of the 4th international conference on Embedded networked sensor systems*, ser. SenSys '06. ACM, 2006, pp. 307–320.

[20] J. Elson and K. Römer, "Wireless sensor networks: a new regime for time synchronization," *SIGCOMM Comput. Commun. Rev.*, vol. 33, pp. 149–154, January 2003.

[21] M. Schuts, F. Zhu, F. Heidarian, and F. W. Vaandrager, "Modelling clock synchronization in the chess gmac wsn protocol," in *Proceedings Workshop on Quantitative Formal Methods: Theory and Applications*, ser. QFM'09, vol. 13, 2010.

[22] I. Rhee, A. Warrier, J. Min, and L. Xu, "Drand: Distributed randomized tdma scheduling for wireless ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 8, pp. 1384–1396, 2009.

[23] S. Cui, R. Madan, A. J. Goldsmith, and S. Lall, "Joint routing, mac, and link layer optimization in sensor networks with energy constraints," in *In Proceedings IEEE ICC 2005*, 2005, pp. 725–729.

[24] S. Ganeriwal, R. Kumar, and M. B. Srivastava, "Timing-sync protocol for sensor networks," in *Proceedings of the 1st international conference on Embedded networked sensor systems*. ACM New York, 2003, pp. 138–149.

[25] J. Levendovszky, K. Tornai, G. Treplán, and A. Oláh, "Novel load balancing algorithms ensuring uniform packet loss probabilities for WSN," in *2011, IEEE 73rd Vehicular Technology Conference, IEEE VTC-Spring*, Accepted in January 2011.

 ISBN: 978-1-61208-126-7

The Impact of Control Information Prioritization on QoS Performance Metrics

Jasmina Baraković
BH Telecom, Join Stock Company
Sarajevo, Bosnia and Herzegovina
e-mail: jasmina.barakovic@bhtelecom.ba

Sabina Baraković
Ministry of Security of Bosnia and Herzegovina
Sarajevo, Bosnia and Herzegovina
e-mail: sabina.barakovic@msb.gov.ba

Himzo Bajrić
BH Telecom, Join Stock Company
Sarajevo, Bosnia and Herzegovina
e-mail: himzo.bajric@bhtelecom.ba

Abstract— **Real-time protocols such as Real-time Transport Protocol (RTP) and its companion protocol Real-time Transport Control Protocol (RTCP) are the solution for customers who have progressively come to desire real-time services. The characteristics of RTP and RTCP packets differ since they carry different content types, i.e., real-time data and control data, respectively. The standard approach assumes that RTP real-time data packets and RTCP control packets are classified into the same media-oriented service class. That might impact the Quality of Service (QoS) for real-time services with regard to network status. Therefore, this paper proposes an idea of QoS enhancement by efficiently transmitting user control information and establishing the proper transmission rate dynamically in dependence upon the status of a network. To accomplish this task, RTCP control packets are classified the Signaling service class, which should be given the absolute preferential treatment over all other User service classes. Results obtained with NCTUns simulator and emulator show a significant impact of control information prioritization on QoS performance metrics and indicate that this approach could be a new starting point for research activities in the future.**

Keywords-control information prioritization; QoS; RTCP; RTP

I. INTRODUCTION

Customers have progressively come to desire real-time services, and the solutions to their desires came in the form of real-time protocols. Using packet switching and the routing of Internet Protocol (IP) packets means that delays are introduced in the access and core networks [1], and hence a data acquired in real-time is transformed undesirably in the networks to non-real-time in the receiver. Real-time protocols are used to correct these delays and the degradation that is introduced during data processing and IP routing in the network. The goal of real-time protocols is to transport, control, and reassemble the information-bearing bits into real-time data for display at the receiver.

The real-time protocols include Real-time Transport Protocol (RTP) [2], which is an Internet Engineering Task Force (IETF) standard. The RTP offers data packet sequencing to enable correct ordering of packets at the receiver because packets can take different routes to their destination. It also provides identification of data source and type of payload, timing and synchronization, monitoring of delivery for diagnosing, reducing transmission problems, and sending the feedback of correct delivery and the quality of data transmission to the sending device. Furthermore, it integrates traffic sources of different types. This is used to merge heterogeneous traffic from multiple transmitting sources into a single flow. Because packets are time stamped at their source and destinations, delays encountered in transit can be estimated and unnecessary jitter can be removed to enable real-time display and enhance Quality of Service (QoS).

This protocol comes together with its companion protocol Real-time Transport Control Protocol (RTCP) [2]. Its function is to provide a means for reporting the performance of data transfer in the network. Roles of RTCP are to expose the states of the client/server to each other so that they understand and exchange the parameters of the communication, and report on the quality of communication. This protocol issues and transmits periodic control packets from participants to all other participants in a session. Two types of RTCP packets are exchanged by the RTCP: the Sender Reports (SRs) and the Receiver Reports (RRs). Fields in the report packets contain descriptions of the state of the session. Compound RTCP packets can be formed by concatenating several report packets and transmitting them as one packet. The QoS can be determined using the parameters in the reports provided by the RTCP [3]. The RTCP reports on the number of packets sent and received since the last report (throughput) and hence the number of lost packets. Because packets are time stamped, the RTCP also provides the round-trip delay, the state of the paths, and of course, the jitter associated with the packets.

These protocols, RTP and RTCP, are designed to be independent of the underlying transport and network layers. In its conventional implementation, RTP does not have a standard Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) port on which it communicates. The only standard that RTP obeys is that UDP communications are done via an even port and the next higher odd port is used for RTCP communications.

Since RTCP packets are sent using a different UDP source and destination port, it is not unlikely that the RTCP packets will receive a different treatment by the network. The standard approach requires that RTCP packets must be marked with the same DiffServ Code Point (DSCP) as the RTP packets in an effect to gain similar treatment from the network as that provided to the RTP packets. However, utilizing the same DSCP for the RTCP packets as that used for the RTP packet does not resolve all problems as follows.

 ISBN: 978-1-61208-126-7

First, RTCP packets vary in size and are generally larger then RTP packets which effects their treatment by a network. Second, RTCP packets are sent at a rate as little as $1/500^{th}$ of the rate that RTP packets are sent which may also affect their treatment by the network. Third, resource reservations made to protect the RTP flows of packets are unlikely to be made to protect the RTCP flow; and if the reservation was made for the RTCP packets, it could fail, and/or be treated differently because of the vastly different traffic profiles. In summary, the QoS statistics determined by RTCP packets may be different than the actual QoS experienced by RTP packets carrying the actual media.

Accordingly, this paper proposes an idea of QoS enhancement by efficiently transmitting user control information and establishing the proper transmission rate dynamically in dependence upon the status of a network. This is attained through the classification of RTCP control packets into the Signaling service class rather than media-oriented service classes. According to our previous work [4], Signaling service class should be given the absolute preferential treatment over all other User service classes. From the RTCP perspective, it is beneficial to receive the absolute preferential treatment over RTP as it provides more accurate statistics for the measurements performed by RTCP, and thereby increase the QoS.

In the remainder of the paper, the related work on QoS evolution and signaling is discussed in Section II. The concept of control information prioritization is described in Section II. It also highlights the difference between the standard and novel approach in terms of control information prioritization. Section III considers the impact of standard and novel approach on QoS performance metrics using simulation methodology. It discusses the obtained results, together with their analysis to show that the conclusions are warranted. Section IV concludes this paper and outlines open issues for future work.

II. RELATED WORK

First studies proposing QoS frameworks for IP networks started to appear within IETF. To support QoS in IP networks, IETF proposed two frameworks. These are Integrated Services (IntServ) [5], based on connection-oriented resource reservation principle and Differentiated Services (DiffServ) [6], based on service differentiation approach. IntServ provides QoS to end hosts by reserving end-to-end resources using the Resource Reservation Protocol (RSVP) [7], when the end hosts signal their QoS needs. DiffServ obviates the need for a resource reservation protocol and offers the benefits of provisioning differentiated services. DiffServ is a starting point to guarantee QoS by providing different service classes, which are configured using specific combination of QoS mechanisms.

Although Multiprotocol Label Switching (MPLS) [8] is not considered as a QoS framework for IP networks, it provides a number of advantageous features to network operators. According to MPLS, data are transmitted along the so-called Label Switched Paths (LSP), which are established using either RSVP modification [9] or specifically developed Label Distribution Protocol (LDP) [10]. Modern QoS-aware networks such as DiffServ, MPLS or DiffServ/MPLS are specifically designed to be flexible enough to reallocate network resources in the best possible way, such that the required performance is provided using minimum amount of resources [11].

To be able to provide QoS when needed, it is necessary not only to deploy proper and effective QoS frameworks, but also to have means to signal to the network entities in charge to set-up QoS the desired level of service. In other words, it is necessary to build a framework that, interworking with a proper protocol, will signal QoS in an efficient and reliable way. The IETF Working Group Next Step in Signaling (NSIS) [12] has been chartered to address these issues, and define the framework for QoS signaling. Moreover, the Telecommunication Industry Association (TIA) has published a standard TIA-1039, QoS Signaling for IP QoS Support, which is involved in solutions to the problem of improving end-to-end QoS performance, e.g., Control for High-Throughput Adaptive Resilient Transport (CHART) system [13]. Another approach related to providing QoS under network congestion is described in [14].

In general, many QoS related projects show that there is an important background of work that aims at achieving QoS provisioning with the different tasks it involves [15]. Neither of referenced approaches discusses the possibility to enhance the QoS for real-time services by efficiently transmitting user control information. Our approach is based on an idea of prioritizing control information transmission in order to improve the QoS performance of real-time services.

III. CONTROL INFORMATION PRIORITIZATION

To accomplish the task of prioritizing control information transmission, DiffServ addresses the clear need for relatively simple and coarse methods of categorizing traffic into various service classes and applying QoS parameters to those classes. Different service classes are constructed using DSCPs, traffic conditioners, Per Hop Behaviors (PHBs), and Active Queue Management (AQM) mechanisms [16].

Though the IETF standards provided a consistent set of PHBs for services marked to specific DSCP values, they never specified which service should be marked to which DSCP value. Cisco led many confusions and disagreements over matching services with standards-defined code points in 2002 to put forward a standards-based marking recommendation in its strategic architectural QoS Baseline document [17]. Eleven service classes that could exist within the enterprise were examined and extensively profiled, then matched to their optimal RFC-defined PHBs. More than four years after Cisco put forward its QoS Baseline document, RFC-4594 was formally accepted as an informational RFC in 2006 [16]. An informational RFC is an industry-recommended best practice. This RFC puts forward twelve service classes and matches these to RFC-defined PHBs.

There are more than a few similarities between Cisco's QoS Baseline and RFC-4594, as there should be, since RFC-4594 is essentially an industry-accepted evolution of Cisco's QoS Baseline. However, there are some differences that

 ISBN: 978-1-61208-126-7

merit attention. Cisco has completed a marking migration for Call Signaling from AF31 to CS3 (as per the original QoS Baseline). The most significant of the differences between Cisco's QoS Baseline and RFC-4594 is the RFC-4594 recommendation to mark Call Signaling to CS5. In summary, the Cisco modified version of RFC-4594 is very similar to RFC-4594, with the one exception of swapping Call Signaling marking and Broadcast Video [17].

Since RFC-4594 is to be viewed as industry best-practice recommendation, enterprises and service providers are encouraged to adopt this recommendation, with the aim of improving QoS consistency, compatibility, and interoperability. However, since it is a set of formal DiffServ QoS configuration best practices, and not a requisite standard, modifications can be made to these recommendations as required by specific needs and constraints. To meet the QoS requirements as defined in International Telecommunication Union – Telecommunication Sector (ITU-T) Recommendation Y.1541, we proposed a modification of these configuration guidelines with regard to Signaling service class [4].

The approach proposed in our previous work is based on configuring Signaling service class by using priority queuing system to give it absolute preferential treatment over all other User service classes. The priority queuing system is a combination of a set of queues and a scheduler that empties them in priority sequence [16]. When asked for a packet, the scheduler inspects the highest priority queue dedicated to Signaling service class and, if there is traffic present, returns a packet from that queue. Failing that, it inspects the next highest priority queue, and so on. A packet assigned to Signaling service class is marked with a new DSCP value, which should be requested from the Internet Assigned Numbers Authority (IANA). This DSCP value should be lower than one used to configure the Network Control service class and higher than one reserved for all User service classes defined in RFC-4594 and RFC-5865 [18].

Though this approach is signaling protocol independent, our previous work has discussed it in the context of Session Initiation Protocol (SIP) [4]. This work considers the previously proposed approach with regard to RTCP.

Since the RTCP packets simply signal information regarding the reception of the RTP packets, we propose to mark them with DSCP value associated to Signaling service class rather than media-oriented service classes. In this manner, it is possible to provide reliable and efficient transmission of user control information in order to increase the QoS by monitoring the media status, reporting on media quality, and taking any necessary corrective actions based on the media status. Thus, the main contribution of this paper lies in highlighting the fact that prioritizing control information transmission improves the QoS performance of real-time services.

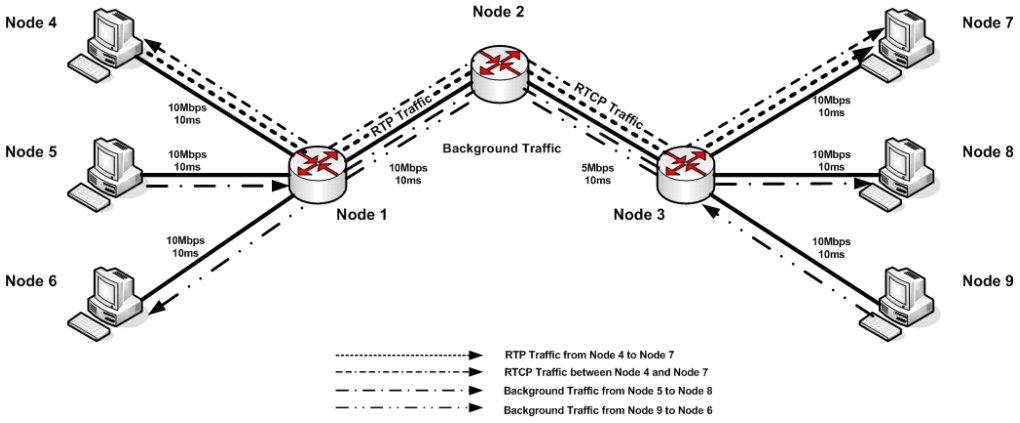
IV. SIMULATION METHODOLOGY

A. Simulation Setup and Environment

In order to investigate the impact of novel approach of control information prioritization on QoS performance metrics, the simulations are performed using NCTUns network simulator and emulator 6.0 [19], because it enables conduction of RTP/RTCP and DiffServ simulations.

The simulations are based on two different simulation scenarios performed on the simple network topology shown on Figure 1. The scenarios differ in control information prioritization. In Scenario 1, RTCP control packets are marked with the same DSCP as RTP packets and classified into the Multimedia Streaming service class. On the other hand, in Scenario 2, RTCP control packets are marked with higher DSCP then one used for RTP packets. The RTCP control packets are classified into Network Control class, since it is the only class in NCTUns that provides the absolute preferential treatment over all other User service classes, as Signaling service class should.

The single DiffServ domain, which constitutes the illustrated network topology, is composed of two boundary routers and one interior router. Boundary routers are responsible for classifying and conditioning traffic. Traffic classification is based on five-tuple (source IP address, destination IP address, source port number, destination port number, protocol). Several parameters are distinguished for this purpose as shown in Table I.



a. RTP (Real-time Transport Protocol); b. RTCP (Real-time Transport Control Protocol).

Figure 1. Network topology.

TABLE I. SIMULATION ENVIRONMENT DETAILS

Node Name	Node Type	Source IP[a] Address	Application	Port	Destination IP[a] Address
Node 1	Boundary Router	NA[b]	NA[b]	NA[b]	NA[b]
Node 2	Interior Router	NA[b]	NA[b]	NA[b]	NA[b]
Node 3	Boundary Router	NA[b]	NA[b]	NA[b]	NA[b]
Node 4	Sending Node	1.0.1.2	adapt_bw	5004, 5005	1.0.6.2
Node 5	Sending Node	1.0.2.2	stg	Default	1.0.7.2
Node 6	Receiving Node	1.0.3.2	rtg	Default	NA[b]
Node 7	Receiving Node	1.0.6.2	rtprecvonly	5004, 5005	NA[b]
Node 8	Receiving Node	1.0.7.2	rtg	Default	NA[b]
Node 9	Sending Node	1.0.8.2	stg	Default	1.0.3.2

a. IP (*Internet Protocol*); b. NA (*Not Applicable*).

Traffic conditioning is based on the token bucket scheme. Since traffic conditioning helps to prevent network congestion, it is set to be disabled in both simulation scenarios. Interior router is responsible for dispatching the incoming packets to the different service class queues for receiving different QoS treatments. The forwarding treatments (i.e., PHBs), which are used for the purpose of investigating the impact of control information prioritization on QoS performance metrics, are listed in Table II.

The links between network nodes are dimensioned to implement simple network configuration with one bottleneck link (Node 2 → Node 3). The configured link capacities, as well as propagation delays are shown in Figure 1. The delay of all links is 10 ms, and links capacity is 10 Mbps except link between Node 2 and Node 3. The bottleneck link capacity is 5 Mbps.

The network is loaded by two types of traffic flows that belong to different service classes. The Standard service class is intended for all background traffic that will receive normal (undifferentiated) forwarding treatment through the network. The Multimedia Streaming service class is used for transport of video traffic using RTP. The RTCP control packets use this service class in Scenario 1, and the Network Control service class in Scenario 2. The Network Control service class is primarily intended for routing and network control. Since this service class has preferential treatment over all other User service classes, it is considered as alternative to the Signaling service class in Scenario 2.

The traffic generators components are stg and rtg, respectively for sending and receiving background traffic. Using UDP greedy mode, stg application on Node 5 generates packets of 1400 bytes during 999 s, whereas stg application on Node 9 generates packets of 1024 bytes during 500 s. The RTP and RTCP traffic is generated using rtprecvonly and adapt_bw application programs, which are implemented on Node 7 and Node 4, respectively. Application rtprecvonly receives RTP and RTCP packets, sends RTCP packets, but does not send RTP packets. Application adapt_bw uses RTCP packets to report the received QoS at the receiver so that sender can dynamically adjust the sending rate of its RTP packets. The RTP packets are received on port 5004, whereas the RTCP packets are received on port 5005. The RTP is used to transport video traffic with characteristics listed in Table III.

The simulations are run for 500 simulations seconds. Due to nearly permanent characteristics of background traffic, this period can be considered sufficient. The RTP/RTCP session active time is from 5th to 500th simulation second. Start time and stop time for background traffic generator implemented on Node 5 is 150th and 350th simulation second, respectively. The activity of background traffic generator implemented on Node 9 is started at 0.5th simulation second and ended at 500th simulation second.

B. Simulation Results and Analysis

Obtained simulation results show significant impact of the novel approach on critical QoS performance metrics in comparison to standard approach, which is particularly obvious during the maximum network load. The standard approach involves marking both RTP and RTCP packets with same DSCP value (i.e., Scenario 1). The novel approach is based on marking RTCP packets with the higher DSCP value than one used for RTP packets (i.e., Scenario 2). The considered performance metrics, which are important for QoS of real-time services, include: throughput, RTP packet loss rate, cumulative number of lost RTP packets, Round-Trip Time (RTT) and jitter for RTP packets.

TABLE II. DISTINGUISHED QOS CONFIGURATION PARAMETERS

QoS[g] Configuration Setup in Boundary Router				
	Source IP[e] Address	***Source Port***	***DSCP[d]***	
			Scenario 1	***Scenario 2***
Rules	1.0.1.2	5004	AF[a]31	AF[a]31
	1.0.1.2	5005	AF[a]31	CS[b]6
	1.0.6.2	*	DF[c]	DF[c]
	1.0.8.2	*	DF[c]	DF[c]
QoS[g] Configuration Setup in Interior Router				
Queues	***PHB[f] Type***	***Rate [Mbps]***	***Queue Length [packets]***	
	DF[c]	7.5	10	
	AF[a]41	2	10	
	CS[b]6	0.5	10	

a. AF (*Assured Forwarding*); b. CS (*Class Service*); c. DF (*Default Forwarding*); d. DSCP (*DiffServ Code Points*); e. IP (*Internet Protocol*); f. PHB (*Per Hop Behavior*); g. QoS (*Quality of Service*).

TABLE III. VIDEO TRAFFIC CHARACTERISTICS

Encoding name	nv [20]
Sampling rate [Hz]	90000
Bits per sample	0.555555
Frame rate [fps]	30

 ISBN: 978-1-61208-126-7

As shown on Figure 2a, the novel approach based on classification of RTCP control packets into the Signaling service class, provides throughput in range from 5 kBps to 20 kBps during the interval of maximum network load. On the other hand, throughput obtained in the standard approach, in which RTP and RTCP packets are classified into the same media-oriented service class, does not exceed 10 kBps during the same period. Therefore, the novel approach results in higher throughput than the standard approach. This is a consequence of the control information prioritization, which makes establishing the proper transmission rate dynamically in dependence upon the status of a network.

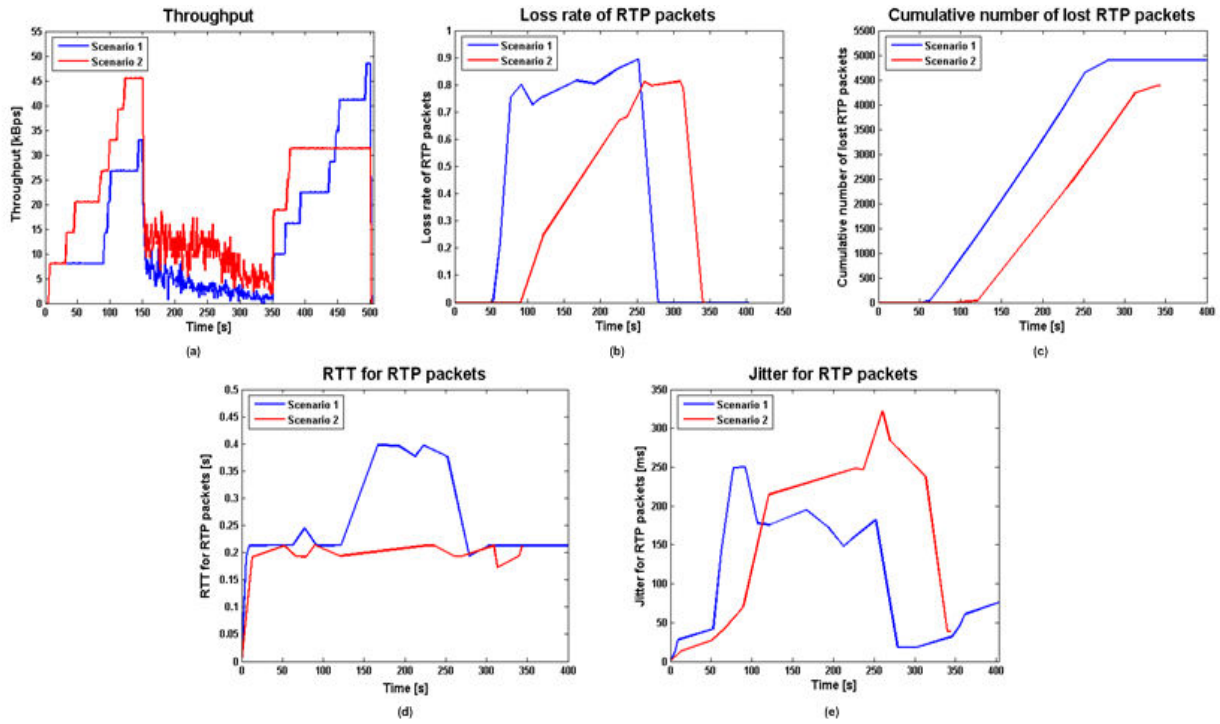
The RTP packet loss rate shown on Figure 2b implicates that the novel approach to the classification of RTCP packets performs better than standard one. In the novel approach, RTCP packets adjust the sending rate of RTP packets more dynamically than in the standard approach. Thereby, packet losses are prevented, which results in lower cumulative number of lost RTP packets (Figure 2c). Accordingly, cumulative number of lost RTP packets equals 4383 and 4892 packets in novel and standard approach, respectively.

Since the novel approach uses RTCP packets for more dynamic adjustment of the sending rate of RTP packets, it also prevents RTP packets to experience long delay caused by network status. Standard approach, on the other hand, does not adjust the sending rate as fast as the novel approach, since the QoS statistics determined by delayed RTCP packets may be different than the actual QoS experienced by RTP packets carrying the media. The lack of the standard approach in comparison with the novel one is shown on Figure 2d. It is noticed that the RTT of RTP packets in standard approach is higher than in novel one during the interval of maximum network load. Moreover, it is kept constant during the simulation, which results in lower jitter than in novel approach (Figure 2e).

V. CONCLUSION AND FUTURE WORK

This paper proposes an idea of improving the QoS performances of real-time services by prioritizing user control information transmission and establishing the proper transmission rate dynamically in dependence upon the status of a network. To accomplish this task, the approach proposed in our previous work is used. According to that approach, Signaling service class is configured using priority queuing system to give it absolute preferential treatment over all other User service classes. Since the proposed approach is signaling protocol independent, this paper considers it with regard to RTCP. In this respect, RTCP packets are classified into the Signaling service class rather than media-oriented service classes in order to transmit the user control information reliably and efficiently.

Obtained results show the significant impact of control information prioritization on QoS performance particularly during the maximum network load. Prioritizing control information transmission provides more accurate statistics for the measurements performed by RTCP, and thereby increases the QoS.



a. RTP (Real-time Transport Protocol); b. RTT (Round-Trip Time).

Figure 2. Simulation results: (a) Throughput; (b) Loss rate of RTP packets; (c) Cumulative number of lost RTP packets; (d) RTT for RTP packets; (e) Jitter for RTP packets.

 ISBN: 978-1-61208-126-7

However, the opportunity for the improvement of this approach may be identified. Encouraged by the positive results, the intention is to elaborate proposed idea and investigate it in detail. That could include the research activities on behavior of various types of multimedia real-time traffic under various network conditions. Additionally, the intention is to emulate the real-time traffic, in order to test the functions and performances of real-world devices and services, and observe how they would perform under various network conditions.

Also, since today's innovation and technology development depends on the users' satisfaction, it is necessary to investigate users' perceived quality of the real-time services with regard to this approach. Being linked to various investigation areas, this idea is hopefully going to become a new starting point for research activities in the future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments that greatly improved the paper.

REFERENCES

[1] S. Baraković, J. Baraković, and H. Bajrić, "QoE Dimensions and QoE Measurement of NGN Services," Proc. Telecommunications Forum (TELFOR 2010), IEEE press, November 2010, pp. 15-18.

[2] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," Technical Report RFC 3550, Internet Engineering Task Force (IETF), July 2003.

[3] J. I. Agbinya, IP Communications and Services for NGN. CRC Press, Taylor and Francis Group, 2010.

[4] J. Barakovic, H. Bajric, and S. Barakovic, "Priority Level Configuration for Signaling Service Class," Proc. IEEE International Conference on Communication Theory, Reliability and Quality of Service (CTRQ 2010), IEEE Press, June 2010, pp. 122-127, doi: 10.1109/CTRQ.2010.28.

[5] B. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an overview," Technical Report RFC 1633, Internet Engineering Task Force (IETF), June 1994.

[6] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," Technical Report RFC 2475, Internet Engineering Task Force (IETF), December 1998.

[7] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification," Technical Report RFC 2205, Internet Engineering Task Force (IETF), September 1997.

[8] E. C. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture," Technical Report RFC 3031, Internet Engineering Task Force (IETF), January 2001.

[9] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels," Technical Report RFC 3209, Internet Engineering Task Force (IETF), December 2001.

[10] L. Andersson, P. Doolan, N. Feldman, A. Fredette, and B. Thomas, "LDP Specification," Technical Report RFC 3036, Internet Engineering Task Force (IETF), January 2001.

[11] J. Barakovic, H. Bajric, and A. Husic, "QoS Design Issues and Traffic Engineering in Next Generation IP/MPLS Network," Proc. IEEE International Conference on Telecommunications (ConTEL 2007), IEEE Press, Zagreb, Croatia, June 2007, pp. 203-210, doi: 10.1109/CONTEL.2007.381873.

[12] X. Fu, H. Schulzrinne, A. Bader, D. Hogrefe, C. Kappler, G. Karagiannis, H. Tschofenig, and S. Van den Bosch, "NSIS: A New Extensible IP Sinaling Protocol Suite," IEEE Communication Magazine 2005, Vol. 43, No. 10, pp. 133-141.

[13] J. Brassil, et al., "The CHART system: a high-performance, fair transport architecture based on explicit-rate signaling," ACM SIGOPS Operating System Review, Vol. 43, No. 1, January 2009, pp. 26-35.

[14] N. Dukkipati, N. McKeown, and A. G. Fraser, "RCP-AC: Congestion Control to Make Flows Complete Quickly in Any Environment," Proc. IEEE International Conference on Computer Communications (INFOCOM 2006), IEEE press, April 2007, pp. 1-5, doi: 10.1109/INFOCOM.2006.18.

[15] Project Economics and Technologies for Inter-Carrier Services (ETICS), "End-to-end service specification template," November 2010. Available at: https://www.ict-etics.eu/, 27.02.2011.

[16] J. Babiarz, K. Chan, and F. Baker, "Configuration Guidelines for DiffServ Service Classes," Technical Report RFC 4594, Internet Engineering Task Force (IETF), August 2006.

[17] V. Joseph and B. Chapman, Deploying QoS for Cisco IP and next generation networks: the definitive guide. Morgan Kaufmann, 2009.

[18] F. Baker, J. Polk, and M. Dolly, "A Differentiated Service Code Point (DSCP) for Capacity-Admitted Traffic," Technical Report RFC 5865, Internet Engineering Task Force (IETF), May 2010.

[19] NCTUns network simulator and emulator web site. Available at: http://nsl10.csie.nctu.edu.tw/nctuns.html, 27.02.2011.

[20] H. Schulzrinne and S. Casner, "RTP Profile for Audio and Video Conferences with Minimal Control," Technical Report RFC 3551, Internet Engineering Task Force (IETF), July 2003.

 ISBN: 978-1-61208-126-7

Quality of Services Assurance for Multimedia Flows based on Content-Aware Networking

Eugen Borcoci, Mihai Stanciu, Dragoş S. Niculescu
Telecommunications Dept.
University POLITEHNICA of Bucharest
Bucharest, Romania
e-mails: {eugenbo, ms, dniculescu}@elcom.pub.ro

George Xilouris
NCSR Demokritos
Institute of Informatics and Telecommunications
Athens, Greece
e-mail: xilouris@iit.demokritos.gr

***Abstract*—This paper proposes a new architectural solution to support Quality of Services (QoS) for real time media flows in a multi-domain system based on new concepts as Content-Aware Networks (CAN) and Network Aware Application (NAA). The system described, based on coupling between network and applications is focused, but not limited to, on multimedia services, with content aware processing in the network, including QoS assurance. The architecture actually parallelizes the Internet in virtual CAN networks, spanning multiple domains and assigning specific quality of services classes to different CANs. This work is a part of the starting effort inside of a new European FP7 ICT research project, ALICANTE.**

Keywords*— *Quality of Service; Content-Aware Networking; Network Aware Applications; Multimedia distribution; Future Internet

I. INTRODUCTION

A new trend in the Future Internet (FI) [1-6] is to increase the coupling between the transport architectural stratum and application layer, the result being that one has content awareness at the network layer and network aware applications at the higher layer. The new concepts are called *Content-Aware Networks* (CAN) and *Network-Aware Applications* (NAA). This approach is supposed to support richer processing of media flow at the network level. The solution is investigated by many groups, given the general accepted vision, stating that [10][11] the FI will be strongly service-content oriented and media oriented. The CANs can be constructed as overlays, on top of traditional IP networks using network virtualization (note that this is seen as a main way to make the Internet more flexible [8][9][12]). CAN routers are optimized for additional tasks (with respect to the traditional ones) such as content/context-based filtering, QoS processing, routing/forwarding, adapting and transforming the packet flows.

This work is a starting activity, performed in the framework of a new European FP7 ICT research project, "Media Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments", ALICANTE, [15] [16][17]. Inter-working environments are defined, to which different cooperating business actors belong: *User Environment (UE)*, to which some End-Users belong; *Service Environment (SE)*, to which Service Providers (SP) and Content Providers (CP) belong; *Network Environment (NE)*, to which the Network Providers (NP) belong. The "*Environment"*, is a generic name for a grouping of functions defined around the same common goal and which possibly vertically span one or more several architectural (sub)layers. By *Service*, if not specified differently, we understand here *high level services*, as seen at the application/service layer. The above environments are actually present in current deployments, but there is insufficient collaboration between them. The neutral network service, considered many years as a basic and good principle (despite that there are large discussions to preserve it or not in FI), proves nowadays to be a weak solution, especially if one considers the new multimedia communications and their increasing importance in the FI.

This paper proposes an enhanced solution for guaranteed QoS assurance in a multi-domain CAN network context. It is organized as follows. Section 2 presents samples of related work. Section 3 summarizes the overall ALICANTE architecture. Section 4 is focused on the CAN solutions for QoS assurance. Section 5 contains some conclusions.

II. RELATED WORK

Nowadays a higher coupling between the Application and Network layers is investigated, targeting to better performance (for multimedia) but without loosing modularity of the architecture. The CAN (i.e., adjusting network layer processing based on limited examination of the nature of the content) and NAA (i.e., processing the content based on limited understanding of the network condition) are studied in the framework of re-thinking the architecture of the FI.

The work [13] considers that CAN/ NAA can offer a way of evolution of networks beyond IP. The capability of content-adaptive network awareness is exploited in [1] for joint optimization of video transmission. The CAN/NAA approach can naturally lead to a user-centric FI and telecommunication services as described in [3][8][9]. The work in [9] discusses the content adaptation issues in the FI

 ISBN: 978-1-61208-126-7

as a component of CAN/NAA approach. The CAN/NAA approach can also offer QoE (Quality of Experience) and QoS capabilities of the future networks, [4][6]. The architecture can be still richer if we add context awareness to content awareness [7]. The CAN approach, on the other side, requires a higher amount of packet header processing in the CAN elements similar to deep packet inspection techniques; therefore, new methods are needed to minimize this processing task. The CAN/NAA approach can also help to solve the current networking problems related to the P2P traffic overload of the global Internet [14]. The application layer traffic optimization (ALTO) problem studied by IETF can be solved by the cooperation between the CAN layer and the upper layer.

III. ALICANTE SYSTEM ARCHITECTURE

The ALICANTE architecture includes network-awareness at the service/application level and content-awareness at the network level. In [15][16][17] the main concepts of ALICANTE are fully introduced.

A flexible business model is defined, composed of traditional SP, CP, NP - Providers and End-Users (EU). A new actor is the CAN Provider (CANP) which is the virtual layer connectivity SP, offering content-aware network services. A new entity is also defined: Home-Box (HB)- which can be partially managed by the SP, the NP, and the end-user. The HB is a physical and logical entity located at end-user's premises and gathering content/context-aware and network-aware information. The HB can be also seen as a CP/SP for other HBs, on behalf of the End User (EU). The HBs cooperate with SPs in order to distribute multimedia services (e.g., IPTV) in different modes (e.g., native multicast or Peer to Peer -P2P).

The architecture is composed of four layers/environments: User Environment, Service Environment plus HB layer, CAN layer and traditional network layer. Two novel virtual layers exist [16][17], (CAN layer for network level packet processing and HB layer for the actual content delivery, in the user proximity) working on top of IP. The virtual *CAN routers* (CANR) perform the CAN processing. They are also called *Media-Aware Network Elements (MANE)* to emphasize their additional capabilities: content and context - awareness, controlled QoS/QoE, security and monitoring features, etc., in cooperation with the other elements of the system.

The SE [17] uses information from the CAN layer to enforce NAA procedures, in addition to user context-aware ones. Per flow adaptation can be deployed at both HB and CAN layers, as additional means for QoS, by making use of scalable media resources.

In the Data Plane, CAN concepts are applied in order to perform network/transport intelligent content-aware processing (QoS, based on provisioning and dynamic adaptation, routing/forwarding, security, etc.). The management and control of the CAN layer is partially distributed; it supports CAN customization as to respond to the upper layer needs, including 1:1, 1:n, and n:m communications, and also allow efficient network resource exploitation. The rich I/F between CAN and the upper layer allows cross layer optimizations interactions, e.g., including offering distance information to HBs to help working in P2P style. At all levels, monitoring is performed in several points of the service distribution chain and regulates a two fold adaptation action, at the virtual HB Layer and at the virtual CAN Layer.

Figure 1 presents the overall architecture and emphasizes the CAN layer and physical perspective of the system. The UMgr, SMgr, CANMgr, NRMgr – are respective managers at user, service, CAN and network levels.

The network infrastructure contains several NP domains (e.g., autonomous systems - AS) and access networks (AN). Each domain has an Intra-domain Network Resource Manager (Intra-NRMgr), as the ultimate authority configuring the network nodes. The CAN layer cooperates with HB and SE layers, seen as users of the CAN services and using network-aware information delivered by the CAN layer. One CAN Manager (CANMgr) exists for each IP domain to assure the consistency of CAN planning, provisioning, advertisement, offering, negotiation installation and exploitation. However, autonomous CAN-like behavior of the MANE nodes can be offered also in a distributed way based on processing individual flows, content-related metadata, and verification of content-related predicates.

The upper layers HB, UE, SE elaborate together network -aware and context-aware applications. The HB layer hosts the service adaptation, service mobility, security, and overall management of services and content. Service Provider(s) may request CAN construction to CANP.

All CAN operations are performed in MANE nodes, installed at the edges of the domains (for scalability). One or several CANs with different capabilities can be defined, installed and offered by each domain. They also can be chained in order to obtain multi-domain spanned CANs. The MANEs perform processing according to the content properties (described by metadata or packet headers or derived by on-fly content-type analysis) also depending on network properties and its current status. The MANE basic set of functions are [16][17]: content-aware intelligent routing, flow adaptation, QoS and resource allocation, filtering and specific security functions, data caching. The CAN/MANE approach offers advantages over conventional routing but raises several challenging open research aspects, given more tasks to be performed by MANE in comparison with traditional routers.

The SP can offer (via CAN layer) QoS guaranteed services, realized at CAN level, by constructing appropriate virtual (unicast or multicast) single or multiple-domain pipes in the network, with or without resource reservations, based on Service Level Agreements (SLA) contracts. Then, as a second level of actions, adaptation actions will be performed, e.g., adapting flows proactively if we have Scalable Video Codec sources.

IV. CAN LAYER AND QOS ASSURANCE

This section presents several aspects related to (V)CANs and QoS; the approach adopted is that one CAN is associated to a given QoS class. Actually one may have several levels of granularity when defining CANs. However, irrespective to

 ISBN: 978-1-61208-126-7

the granularity, the main common idea is preserved: that the CAN layer offers to the SP "Parallel Internets" specialized at different types of applications content. We consider below the definitions related to QoS classes (QC), in order to establish the relationship between CANs and QoS classes. These classes should be combined when a CAN (represented for instance by a "QoS plane") is spanned over multiple domains.

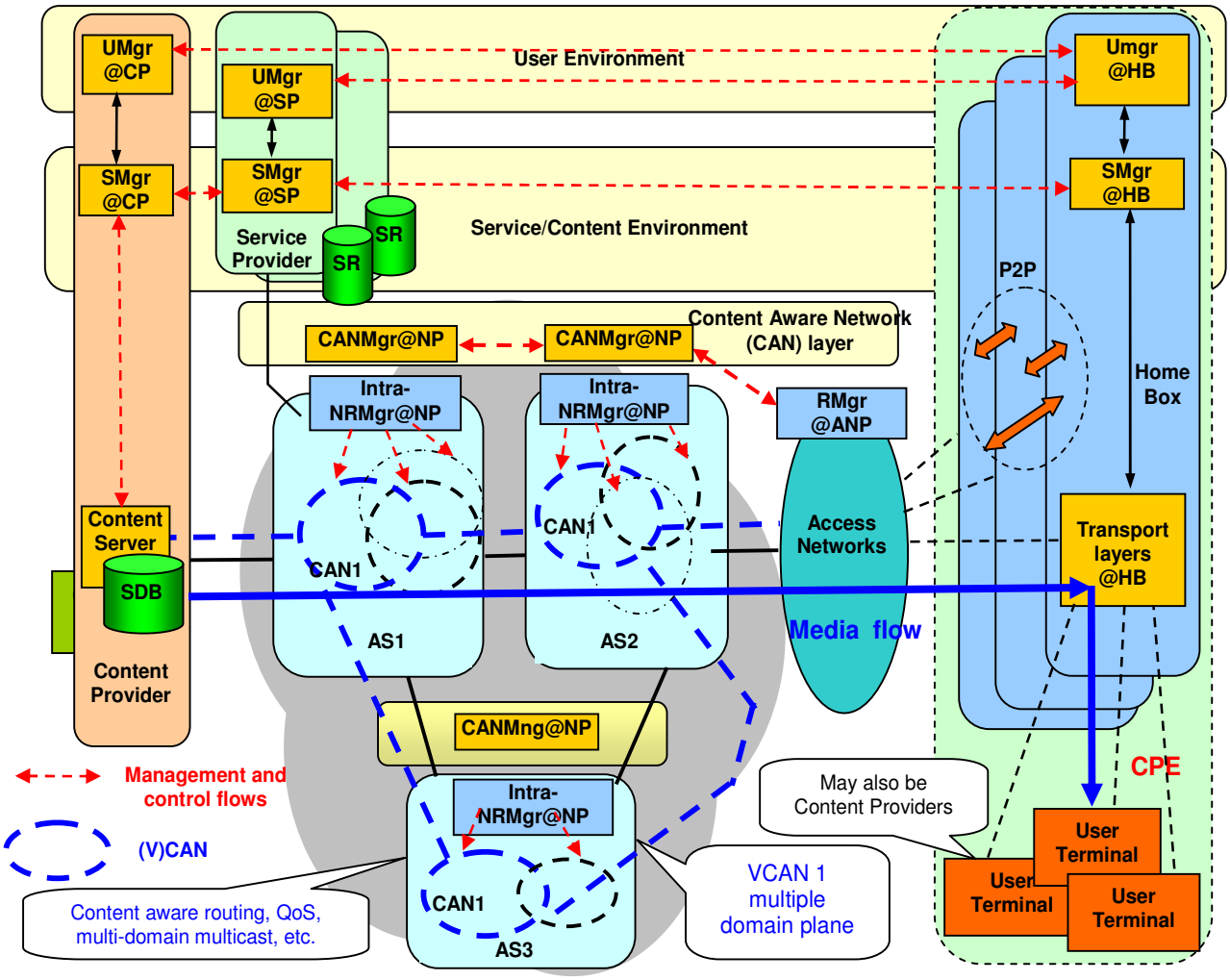


Figure 1. The ALICANTE Overall Architecture

A. *QoS Classes Definitions*

A (well known) QoS class is a QoS transfer capability [18][19] represented by a set of attribute-value pairs, expressing various packet transfer performance parameters such as one-way transit delay, packet loss, jitter. A provider domain's supported QCs can be divided [18][19][22] into *local* QoS classes (l-QCs) and *extended* QoS classes (e-QCs), to allow us to capture the notion of QoS capabilities across domains. From a service offering perspective, QoS classes correspond to the performance (transfer quality) guarantees expressed in contracts as SLSs. From a service provisioning perspective, the QCs classes split the network QoS space into a number of distinct classes, and hence set the traffic-related objectives of traffic engineering functions:

- QoS class (QC): a basic network-wide QoS transfer capability of a single provider's domain. It is defined (in DiffServ but not only) as a set of parameters expressed in terms of {Delay, Jitter, Latency}.
- Local QC (l-QC): a QC that spans a single AS. This is a notion similar to Per Domain Behavior (PDB) – in DiffServ technology.
- Extended QC (e -QC): a QC that spans several ASes. It consists of an ordered set of l-QCs. The topological scope of an e-QC usually extends outside the boundaries of the local domain.

An abstract and flexible definition is the Meta-QoS-Class, MQC [18]. It captures a common set of QoS ranges of

 ISBN: 978-1-61208-126-7

parameters spanning several domains. It relies on a worldwide common understanding of application QoS needs (e.g., VoD service flows need similar QoS characteristics, whatever the transited AS is). The MQC concept offers the advantage that the existence of "international" well known classes greatly simplifies the inter-domain signaling in the sequence of action to establish domain peering in the multi-domain context.

A Meta-QoS-Class can be defined with the following attributes, such as [18]: a list of services (e.g., VoIP) for which the MQC is particularly suited; boundaries for QoS performance attribute (one-way transit delay, one-way transit variation delay –jitter, loss rate); constraints on the ratio: resource for the class - to traffic for the class.

The attributes could depend on AS diameter (e.g., a longer delay could exist in a large AS, and performance attributes can be weighed in order to prioritize those ones to which the service is more sensitive). A given MQC in an AS followed by the same MQC in the next AS should equal the same MQC (invariance).

There is a flexible relationship between a MQC and l-QC of a domain (i.e., not one-to-one relationship) :

- several MQCs (defined by different values of DiffServ DSCP codes), can be mapped onto the same l-QC,
- vice-versa: one can define several l-QC which belong to the same MQC (this means that any such l-QC could be composed with any l-QC of a neighbor domain if the latter l-QC belongs to the same MQC). If for the same service (e.g., VoD) several qualities are wanted, then a hierarchy of MQCs should be defined.

The MQC concept is useful in practice only if a limited set of Meta-QoS-Classes are defined. Each AS classifies its own l-QCs based on Meta-QoS-Class. An l-QC from an AS can be bound only with a neighbor l-QC that refers to the same Meta-QoS-Class. A Meta-QoS-Class typically bears properties relevant to the crossing of one and only one AS. However, this notion can be extended in a straightforward manner to the crossing of several ASes, as long as we consider the set of ASes as a single super-AS.

B. CANs and Network Planes

In ALICANTE, a concept of Parallel Internets (PI) [20] will be adopted, but modified and enriched with content awareness. A PI enables end-to-end service differentiation across multiple administrative domains. The PIs can coexist, as parallel logical networks composed of interconnected, per-domain, Network Planes (NPl). In ALICANTE, generally a one-to-one mapping between a CAN and NPl will exist.

A CAN Network is an overlay seen as a network plane (NPl) + content awareness. Specialization of CANs may exist in terms of QoS level of guarantees (weaker or stronger), QoS granularity, content adaptation procedures, degree of security, etc.

A given NPl is defined to transport traffic flows from services with common connectivity requirements. The traffic delivered within each NPl receives differentiated treatment both in terms of forwarding and routing, so that service differentiation across NPls is enabled in terms of edge-to-edge QoS, availability and also resilience.

A given NPl/CAN can be realized by the CANP, by combining several processes [17], while being possible to choose different solutions concerning some "dimensions":

- Routing: different paths can be implemented for individual CANs/NPls in order to support heterogeneous service requirements. Routing differentiation can be realized at several levels, e.g., one can define dedicated topologies to get several routing adjacencies towards the destination; dedicated paths selection to achieve multiple path (based on different routing metrics in different NPls).
- Data plane forwarding and packet processing: different classification, metering and drop policies, different packet scheduling behavior by configuring different policies in a common scheduler, assigning dedicated scheduling resources, etc.
- Resource management: Data packets can be treated differentially in terms of policing, shaping, degree of multiplexing, over-provisioning factor, queuing, etc.

The CAN granularities from the QoS point of view can be as follows (classified from a lower degree to higher degree):

C. Multi-domain CANs based on Meta-QoS-classes

This is the simplest implementation of CANs. Each (V)CAN, spanning one or more domains, has associated a given MQC. The resulting Internet appears as a set of PIs or equivalently CANs, or MQC planes. Each Internet consists in all the l-QCs bound in the name of the same MQC. If an l-QC maps several Meta-QoS-Classes, then it belongs to several Internets. The SP can define several CANs represented as PIs. The metadata inserted in the data packets allow ingress MANE to select the VCAN that is the closest to its needs (the "best match" principle is preserved), as long as there is currently a path available for the destination.

In a MQC plane, all paths are considered (to a reasonable extent) as equal. Therefore, the problem of path selection amounts to find one best path, for the selected MQC plane. This principle is similar to the traditional IP routing approach. So, for the inter-domain part, one can rely on a BGP-like protocol doing the path inter-domain selection process.

The DiffServ concept of Per-Domain Behavior (PDB) is not identical with the MQC concept. The two concepts both specify some QoS performance values, but they differ in their purposes. The PDB objective is to help implement QoS capabilities within a network, while the MQC definition objective is to help agreement negotiation between CANPs/SPs. Actually, a PDB is closer to an l-QC than to a MQC. The main advantage of the MQC concept is that it simplifies the horizontal negotiations between CAN Managers to chain the single-domain CANs into multi-domain CANs.

Examples of CANs associated each to one of the four MQCs can be: *Premium MQC*; *Gold MQC* for TCP-friendly (elastic) traffic; *Gold MQC* for non TCP-friendly (non-

 ISBN: 978-1-61208-126-7

elastic) traffic; *Best effort MQC*. Examples of basic groupings can be: the overall network organized as an Internet with four Meta-QoS-Classes and an Internet with only the last three Meta-QoS-Classes.

For each MQC one should define parameters. An example for a Premium MQC could be: *Usage:* real time flows, needing constant bandwidth guarantees; *Performance:* Delay, Jitter, Loss – very low (qualitative); *Constraints:* admission control will be applied and possibly shaping to enforce the resource requirement; *Resources:* on each output interface, the traffic for the class is always much smaller than the bandwidth reserved for the class (EF based). The resources must always absorb the traffic with no loss, even with burst aggregates.

Mapping of high level services on such CANs has a low granularity. For instance, if one adopts the TISPAN taxonomy [21], the TS 23.107 document identifies four QoS classes: conversational class, streaming class, interactive class, and background best effort. In such a case VOIP and Video conference will be included in the first class, given their interactivity.

The ALICANTE mapping of service flows on CANs based on MQC approach could be done by defining four type of CANs each one having its well-known MQC. Therefore the four CANs expose decreasing QoS capabilities irrespective to what the type of the service is (video, audio, data). The content awareness of the system means to classify the flows according to their QoE class (gold, silver, etc.) and assign them to be transported over the appropriate CAN/MQC-plane.

D. Multi-domain CANs based on local QC composition

This case is more complex than the previous one. Each domain may have its local QoS classes. Several Local QCs (l-QC) can be combined to form an Extended QC (e-QC). Composition rules for QoS classes should be defined. The granularity is greater than in the MCQ approach, in the sense that a greater number of parallel CANs can be defined. Still, a CAN plane offers a single resultant QoS class.

E. Multi-domain hierarchical CANs based on local QC composition

This case is the most efficient but also the most complex. Each domain may have its local QoS classes. Several Local QCs (l-QC) can be combined to form an Extended QC (e-QC). The difference from the previous solution is that inside each CAN several QOS classes are defined corresponding to platinum, gold, silver, etc. In such a case, the mapping between service flows at SP level and CANs can be done per type of the service: VoD, VoIP, Video-conference, etc. – if SP wants to do it. Additionally to the required QC, a priority indicator can be considered (this is a figure established by the SP), indicating the current priority seen from the business point of view by the SP; e.g., for an emergency situation, a video service flow may have a greater priority than for a high definition entertainment video flow. The granularity is greater than in the MCQ approach, in the sense that a greater number of parallel CANs can be defined. Still, a CAN plane offers a single resultant QoS class.

F. CAN provisioning and content aware processing of service flows

Figure 2 shows an example of two provisioned multi-domain CANs, spanning respectively: CAN1: AS1, AS2; CAN2: AS1, AS2, AS3; based on the implementation in AS1, 2, 3 of the MQC1, 2 as shown.

It was supposed that a functional block CAN-NRMgr exists at SP to initiate the CAN construction conforming to the needs of SP (e.g., based on forecast traffic data). It is shown in a simplified way how the CANs are realized and how the QoS based on content classification is realized in the first MANE of AS1. The following actions are performed:

- CAN 1,2, have been requested (negotiation) by CAN NRMgr@SP, to CANMgr@AS1. The topological data and QoS needs are delivered by the CAN NRMgr@SP (action 1 in Figure 2).
- CANMgr@AS1 negotiates resources with IntraNRM@AS1 (action 2).
- Multiple domain CANs are needed, so CANMgr@AS1 negotiates SLAs (actions 3.1, 3.2) with CANMgr@AS2, CANMgr@AS3 (hub inter-domain peering model is supposed here). The inter-CANMgr negotiations are not visible at SP level.
- CAN1, 2 are installed in the network at SM@SP request (actions 4, 4.1 and 4.2)
- MANE1 is instructed how to classify the data packets, based on information as: VCAN_Ids, Content description metadata, headers to analyze, QoS class information, policies, PHB – behavior rules, etc. obtained from CAN NRMgr@SP via CANMgr@AS1 and IntraNRM@AS1. Also, the output part of the MANE1 is configured for queuing and scheduling as to realize the QoS classes.
- Service Management at SP instructs the SP/CP server how to mark the data packets. The information to be used in content aware classification can be composed of high level headers (e.g., RTP); content description metadata (including optionally an explicit VCAN_Id to simplify the MANE analysis task

The implementation of the QoS classes in the network can be done based on known QoS technologies like MPLS + DiffServ. The mapping of the (V)CAN onto actual paths is performed by the Intra NRMgr. Applying the above procedures, the data flows are classified by the MANE, assigned to the appropriate CAN and processed according to the QoS class associated to that CAN.

A data packet is analyzed by the classifier and assigned to one of the CANs, depending on: MANE classification information and policies; data packet high-level protocol headers and/or metadata and/or VCAN_Id contained in the packet. Consequently, the packet is forwarded to the appropriate logical CAN for further processing.

V. CONCLUSIONS AND FUTURE WORK

The paper proposed an architectural solution supporting QoS in a virtualized network environment, based on content awareness at network level, in the framework of a complex

 ISBN: 978-1-61208-126-7

system for media distribution. The approach is to map the QoS classes on virtual data CANs, thus obtaining several parallel QoS planes. The system can be incrementally built by enhancing the edge routers functionalities with content awareness features. Further work is going on to fully validate the concept and then to design and implement the system in the framework of the FP7 research project ALICANTE.

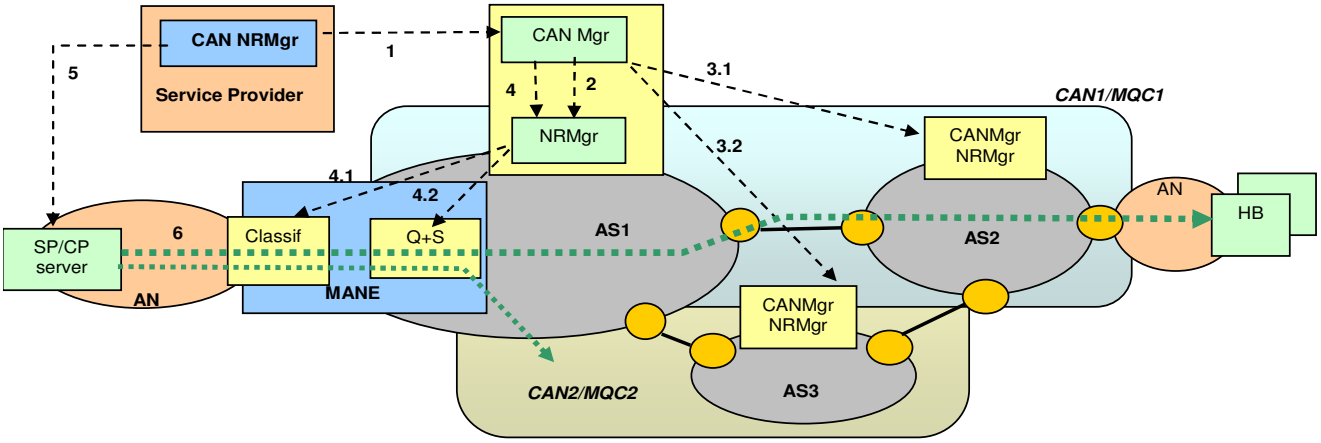


Figure 2. Example of multi-domain CANs based on MQCs

ACKNOWLEDGMENTS

This work was supported partially by the EC in the context of the ALICANTE project (FP7-ICT-248652) and partially by the project POSDRU/89/1.5/S/62557.

REFERENCES

[1] Martini, M.G., Mazzotti, M., Lamy-Bergot, C., Huusko, J., and Amon, P., "Content Adaptive Network Aware Joint Optimization of Wireless Video Transmission", IEEE Communications Magazine, vol. 45, no. 1, Jan. 2007, pp. 84-90.

[2] Baladrón, C., "User-Centric Future Internet and Telecommunication Services", in: G. Tselentis, et al. (eds.), Towards the Future Internet, IOS Press, 2009, pp. 217-226.

[3] Zahariadis, T., Lamy-Bergot, C., Schierl, T., Grüneberg, K., Celetto, L., and Timmerer, C., "Content Adaptation Issues in the Future Internet", in: G. Tselentis, et al. (eds.), Towards the Future Internet, IOS Press, 2009, pp.283-292.

[4] Liberal, F., Fajardo, J.O., and Koumaras, H., "QoE and *-awareness in the Future Internet", in: G. Tselentis, et al. (eds.), Towards the Future Internet, IOS Press, 2009, pp. 293-302.

[5] Baker, N., "Context-Aware Systems and Implications for Future Internet", in: G. Tselentis et. al. (eds.), Towards the Future Internet, IOS Press, 2009, pp. 335-344.

[6] Huszák, Á. and Imre, S., "Content-aware Interface Selection Method for Multi-Path Video Streaming in Best-effort Networks", Proc. of 16th International Conference on Telecommunications, Marrakech, Morocco, Jul. 2009, pp. 196-201.

[7] Kodeswaran, S. B. and Joshi A., "Content and Context Aware Networking Using Semantic Tagging", Proc. of 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, Georgia, USA, Apr. 2006, pp. 67–77.

[8] Anderson, T., Peterson, L., Shenker, S., and Turner, J., "Overcoming the Internet Impasse through Virtualization", Computer, vol. 38, no. 4, Apr. 2005, pp. 34–41.

[9] Turner, J. S. and Taylor, D. E., (2005), "Diversifying the Internet", Proc. GLOBECOM '05, vol. 2, St. Louis, USA, Dec. 2005, pp.760-765

[10] Schönwälder, J., Fouquet, M., Rodosek, G.D., and Hochstatter, I.C., "Future Internet = Content + Services + Management", IEEE Communications Magazine, vol. 47, no. 7, Jul. 2009, pp. 27-33.

[11] Networked European Software and Sevices Initiative (NESSI) Strategic Research Agenda, Vol. 3. FP7-2.exec, NESSI Roadmap, Feb. 2008.

[12] Chowdhury, N. M. and Boutaba, R., "Network Virtualization: State of the Art and Research Challenges", IEEE Communications Magazine, vol. 47, no.7, Jul. 2009, pp. 20-26.

[13] Kourlas, T., "The Evolution of Networks beyond IP", IEC Newsletter, vol. 1, Mar. 2007. Available at http://www.iec.org/newsletter/march07_1/broadband_1.html (last accessed: Mar. 2010).

[14] Aggarwal, V. and Feldmann, A., "Can ISPs and P2P Users Cooperate for Improved Performance?", ACM SIGCOMM Computer Communication Review, vol. 37, no. 3, Jul. 2007, pp. 29-40.

[15] FP7 ICT project, "MediA Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments", ALICANTE, No248652, http://www.ict-alicante.eu/ (last accessed: Mar. 2010).

[16] Borcoci, E., Negru, D. and Timmerer, C., "A Novel Architecture for Multimedia Distribution based on Content-Aware Networking" Proc. of. CTRQ 2010, Athens, June 2010, pp. 162-168

[17] ALICANTE, Deliverable D2.1, ALICANTE Overall System and Components Definition and Specifications, http://www.ict-alicante.eu, Sept. 2010

[18] Levis, P., Boucadair, M., Morrand, P., and Trimitzios, P., "The Meta-QoS-Class Concept: a Step Towards Global QoS Interdomain Services", Proc. of IEEE SoftCOM, Oct. 2004.

[19] Howarth, M.P. et al., "Provisioning for Interdomain Quality of Service: the MESCAL Approach", IEEE Communications Magazine, June 2005, pp. 129-137

[20] Boucadair, M. et al., "A Framework for End-to-End Service Differentiation: Network Planes and Parallel Internets", IEEE Communications Magazine, Sept. 2007, pp. 134-143

[21] 3GPP TS 23.107 version 9.0.0 Release 9, 2010

[22] MESCAL D1.2: "Initial Specification of Protocols and Algorithms for Inter-domain SLS Management and Traffic Engineering for QoS-based IP Service Delivery and their Test Requirements", January 2004, www.mescal.org (last accessed: Dec. 2010).

 ISBN: 978-1-61208-126-7

Estimation of Quality of Experience in 3G Networks with the Mahalanobis Distance

Karel De Vogeleer, Selim Ickin, Markus Fiedler, David Erman, and Adrian Popescu
School of Computing
Blekinge Institute of Technology
371 79 Karlskrona, Sweden
{kdv,sic,mfi,der,apo}@bth.se

***Abstract*—Quality of Experience is a parameter used to express the relationship between Quality of Service and the satisfaction of network service subscribers. The modeling of Quality of Experience demands for solving a multidimensional problem. In this paper, we present a Quality of Experience analysis of streaming videos. Related to this, we show that we can reduce the dimensions of the Quality of Experience modeling with the help of Principle Component Analysis techniques. We demonstrate that for our data set the Zero Throughput Time and the Packet Delay Variation are enough to get a picture of the state of the network. We further calculate the *Mahalanobis distance* to analyze the outliers in the data set. We illustrate that for our data set the 97.5 % quantile for the *Mahalanobis distance* is a good threshold that indicates low user perception. We also advocate the use of robust statistics in the analysis of Quality of Experience as we are dealing with contaminated data sets.**

***Index Terms*—Mahalanobis Distance; Quality of Experience; video streaming; robust statistics; 3G network measurements;**

I. Introduction

Quality of Experience (QoE) modeling is an important aspect for understanding the impact of Quality of Service (QoS) on service subscriber's satisfactions. Current state of the art QoE models are producing results that are approximately able to forecast the QoE for mobile streaming videos. Yet, a single highly effective metric for predicting the QoE in mobile networks valid in any context is to be established. This paper's aim is to contribute to a better understanding of the QoE and QoS relationship in 3rd Generation (3G) networks.

Linear weighting of QoS metrics has been popular in measuring the QoE [1], [2], [3]. The input of these algorithms are numerous QoS metrics that are collected during runtime and are linearly weighted against each other to produce a QoE estimate. The objective is to find a single metric that can be used to estimate the QoE. For this purpose, we introduce a new metric to QoE modeling, the *Mahalanobis distance* [4]. The Mahalanobis distance is a distance metric that expresses the distance of a measurement point to the center of a data set, taking into account the correlation of data set. We employ the Mahalanobis distance to compute a single value over multiple QoS measurements that tells us how a particular data point is related to the average state, i.e., the center of the data set, of the network. We show that the Mahalanobis distance is correlated to QoE even when computed over a subset of the measured QoS metrics.

To analyze the efficiency of the Mahalanobis distance in estimating the QoE we conducted a set of experiments where users watched a video that was streamed over a 3G network to a mobile device. The users rated the video on a scale from 1 to 5 with regards to the image quality, as specified by the ITU-T recommendation [5]. A different class of QoE research employs traditional point-based metrics, e.g., Peak Signal to Noise Ratio (PSNR), Moving Pictures Quality Metric (MPQM), or Mean Square Error (MSE), to assess the QoE of video images [6]. The PSNR is computed by comparing the original image before streaming and the actual image that was displayed on the end user's device. In contrast to PSNR, where computer algorithms are used to evaluate the satisfaction, we used humans to assess the video quality. This yields more realistic User Ratings (URs) but also introduces more noise to the measurement of the UR. Sources of noise originate from the human behavior, e.g., delay in rating, human forget factor, inaccurate rating.

We also demonstrate that the modeling can be simplified by only using a subset of the available QoS metrics without loosing much accuracy. We use Principle Component Analysis (PCA) techniques to reduce the dimensionality of the QoS metrics. Reducing the dimensionally of the QoS metrics simplifies the acquisition and reduces the computational efforts to prognosticate the QoE.

Moreover, we advocate the use of robust statistical methods. In our case the QoS data shows contaminated distributions. The contaminated part is of particular interest as we show that the low QoE ratings correspond to the contaminated part of the data set. Classical statistics fail to identify this contamination whereas robust statistics are designed to deal with anomalies in data sets. By means of an example we demonstrate that the Mahalanobis distance produces more accurate results with robust methods compared to classical methods.

The paper is as follows; in section two, we elaborate on the mathematical concepts that we use during the analysis of our results, followed by an overview of the used QoS metrics. Section three applies the Mahalanobis distance to our data set. We discuss the distribution of the Mahalanobis distance, and identify and model the outliers of our QoS metrics. Section four briefly explains the practical use of the Mahalanobis distance in a real-time environment and we conclude in section five with the conclusion and future work.

 ISBN: 978-1-61208-126-7

II. BACKGROUND

In this section, we provide an overview of the *Mahalanobis distance* and provide an insight in the QoS metrics used for the assessment of the QoE.

A. Mahalanobis Distance

Mahalanobis distance is a distance metric that expresses the distance of a measurement point to the center of its data set taking into account the data set's correlation. The *Mahalanobis distance* differs from the *Euclidian distance* in the use of the correlation between the components of the data set.

The *Mahalanobis distance* is formally defined as

$$D_M(\mathrm{x}) = \sqrt{(\mathrm{x}-\mu)^T S^{-1} (\mathrm{x}-\mu)} \tag{1}$$

where $\mathrm{x} = (x_1, x_2, \ldots, x_N)$ is a multivariate vector containing a set of measurements, $\mu = (\mu_1, \mu_2, \ldots, \mu_N)$ is the center of the data set and S is the covariance matrix of the data set. It can be shown that if a data set is multinormal distributed then the Mahalanobis distances of the data set are distributed approximately as a chi-square distribution:

$$D_M^2 \sim \mathcal{X}_p^2, \tag{2}$$

with p degrees of freedom [7].

We further observe that our data set is contaminated. In such situations classical statistics yield biased estimates. To obtain more accurate estimates we employ the Minimum Covariance Determinant (MCD) algorithm to calculate the covariance matrix. The fast MCD algorithm is a highly robust estimator of multivariate location and scatter [8], [9]. In our paper, we used the `R` implementation of MCD with `alpha` equal to 0.75 and default remaining parameters. This corresponds to a breakdown point of 25 %, which ensures reasonable efficiency and high robustness against outliers [10].

B. QoS Metrics

During our experiments we assessed the following five QoS metrics: Packet Delay Variation (S_D), Packet Rate (PR), Packet Loss Rate (PL), Clumping Rate (CL), and Zero Throughput Time (T_Z). S_D is calculated as

$$S_D = \sqrt{\frac{1}{N-1}\left[\sum_{n=1}^{N}(D_n^2) - N\bar{D}^2\right]}, \tag{3}$$

where N is the number of received packets per interval, D_n the one-way-delay of packet n, and $\bar{D}$ the average one-way-delay. CL is an indicator that detects clumped packets sequences. T_Z is the duration of no observance of throughput. Additionally we count the number of packets, PR, and the packet loss, PL, per second. The T_Z, S_D and CL are different metrics that describe the variation in one-way-delays of packets. More detailed descriptions of the presented QoS metrics can be found in [11].

Before computing the Mahalanobis distances over the QoS metrics we applied an Exponentially Weighted Moving Average (EWMA) to the data set. The EWMA is used to address the human forgetfulness factor and the delay in rating of the objects under investigation. EWMA is defined as

$$y(j) = (1-\alpha)\, x(j) + \alpha\, y(j-1), \tag{4}$$

where α is the smoothing factor. In previous research $\alpha = 0.75$ yielded the most satisfying results [12]. Note that the EWMA needs a warm-up time to become effective.

III. TESTBED SETUP

We set up a testbed to record URs together with QoS metrics while streaming videos. We streamed a video from a Darwin Streaming Server (DSS) [13] to a HTC Dream over the Real Time Streaming Protocol (RTSP) [14]. The users that participated in the experiment watched a video alone in a darkened room and rated the quality of the video image whenever he or she felt was appropriate by pushing one of the five buttons on the screen.

The DSS is running on a Linux (2.6.27) Ubuntu 9.04 machine and streams a video MPEG-4 compressed with dimensions 176×144 pixels, 24 kHz AAC stereo sound, 23.97 fps, and streamed at a rate of 325 kbps. The HTC Dream runs a custom designed video streaming application on Android 1.5.

More detailed description of the testbed and how the measurements are acquired refer to [15].

We collected data during forty minutes of video streaming over a 3G network. A total number of 106 URs were recorded. The ratings in relation to the five measured QoS metrics after EWMA are analyzed.

IV. MEASUREMENT ANALYSIS

We start the analysis of the QoS starts with the identification of the metrics that are most influential in the data set and correlate most to the UR. Then we proceed to compute the Mahalanobis distance over the selected metrics and analyze its distribution. With this information we then try to model the QoS metrics of concern and attempt to optimize the QoE model.

A. Data set reduction

We first look at the pairwise Pearson's correlation coefficient $\rho_s(QoS_i, QoS_j)$ and place the values in the correlation matrix S_P. The correlation matrix (S_P) is presented in TABLE I. The correlation between the QoS metrics over the whole data set is given. The correlation between QoS metrics and URs is computed only over the rated measurement points. We identify

TABLE I
THE CORRELATION MATRIX (S_P) OF THE MEASURED QOS METRICS AND THE UR COMPUTED WITH PEARSON'S CORRELATION COEFFICIENT (ρ).

	S_D	RP	PL	CL	T_Z	UR
S_D	**1.000**	-0.420	0.130	-0.158	-0.123	0.383
RP	-0.420	**1.000**	0.053	0.494	0.251	-0.442
PL	0.130	0.053	**1.000**	0.169	0.155	-0.382
CL	-0.158	0.494	0.169	**1.000**	0.698	-0.645
T_Z	-0.123	0.251	0.155	0.698	**1.000**	-0.662
UR	0.383	-0.442	-0.382	-0.645	-0.662	**1.000**

 ISBN: 978-1-61208-126-7

that some QoS metrics correlate moderately ~ 0.69 while others show less correlation ~ 0.42, or very low correlation. In particular the PL seems to be a very poor estimate of the UR, given its low correlation. As other QoS metrics show better correlation among each other and the UR, a Principle Component Analysis (PCA) is an appropriate technique to see if the dimensionality of the data set can be reduced without the loss of much information.

We apply the Robust Principal Component Analysis (ROBPCA) [16] to our QoS data set with $n = 2245$ measurement vectors and $p = 5$ dimensions. The loadings of the Principle Components (PCs) are shown in TABLE II together with their eigenvalues. The PCs can be seen as the spectrum of the data

TABLE II
THE DECOMPOSITION OF THE QOS DATA IN ITS PRINCIPLE COMPONENTS BY THE PCA TECHNIQUE. THE LOADINGS OF THE QOS METRICS ARE GIVEN PER PC TOGETHER WITH ITS EIGENVALUE λ.

	PC_1	PC_2	PC_3	PC_4	PC_5
λ	185.668	50.747	23.900	0.569	0.000
φ	0.712	0.906	0.998	1.000	1.000
S_D	-0.397	0.256	-0.881	0.022	0
PR	0.151	-0.928	-0.336	0.064	0
PL	0.000	0.000	0.000	0.000	1
CL	0.001	-0.054	-0.041	-0.998	0
T_Z	-0.905	-0.267	0.330	0.000	0

set where λ, the eigenvalues of the covariance matrix, are proportional to the importance of the PC. The MCD algorithm was also used to compute the covariance matrix for the ROBPCA. We used the `R` implementation with `alpha` set to 0.75. Reducing the dimension is achieved by disregarding PCs. A common selection criterion for PCs is

$$\sum_{j=1}^{k} \tilde{l}_j \Big/ \sum_{j=1}^{r} \tilde{l}_j > x, \tag{5}$$

where $\tilde{l}_j$ is the jth eigenvalue of the covariance matrix, $\tilde{l}_1 \geq \tilde{l}_2 \geq \ldots \geq \tilde{l}_r$ with $r = \text{rank}(S)$, and x is proportional to the compression of the data set after PC reduction. φ in TABLE II is the covariance matrix cumulative sum of the relative eigenvalues of the data set. By keeping PC_1 and PC_2 from TABLE II we obtain $x \approx 90\,\%$. The first two PCs are loaded, in order of significance, by T_Z, S_D, PR, CL, and PL. T_Z seems to be the main contributor in PC_1 and PR in PC_2. S_D is the second largest contributor to variation in both PCs and overall more influential than PR. Hence, the data set can be described by T_Z and S_D with regards to its variation, given that the loadings of T_Z and S_D are approximately $90\,\%$ and $35\,\%$ of PC_1 and PC_2, with weights 0.712 and 0.195, respectively.

We observe that the n dimensions of the QoS data set can be reduced while approximately maintaining variability. We show that retaining only T_Z and S_D after dimension reduction is appropriate. By using PCA techniques we concluded that mainly T_Z but also S_D are the QoS metrics that best describe the data set with regards to its variability.

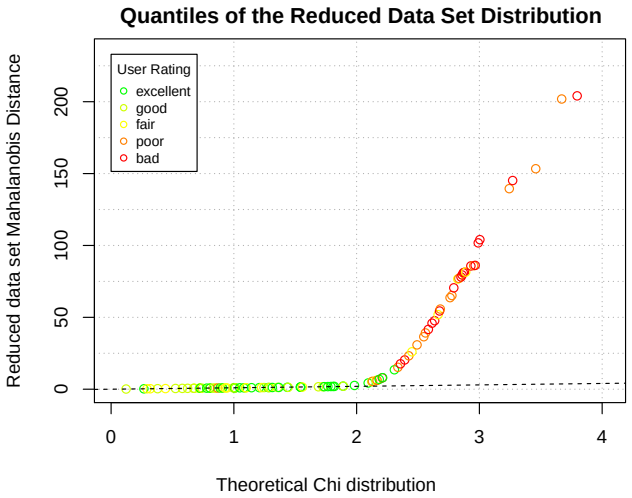


Fig. 1. The Mahalanobis distance Q-Q plot of the reduced data set (T_Z and S_D) against a $\mathcal{X}_{p=2}$ distribution. Only the rated measurements are shown and colored by their User Rating (UR).

B. Mahalanobis distance quantile analysis

We model the QoE with a reduced data set retaining the two most influential QoS metrics on the variance of the data set, i.e., T_Z and S_D. The Mahalanobis distance distribution computed over T_Z and S_D of the whole data set is shown in Figure 1. This figure shows a Q-Q plot where the rated data points are colored to their UR. *Red* is a bad rating, whereas *green* is an excellent rating. The Mahalanobis distance distribution is plotted against a theoretical $\mathcal{X}$ distribution with two degrees of freedom.

The Q-Q plot of the reduced data set against the theoretical $\mathcal{X}_2$ disribution starts of as it seems linearly but it diverges at a given point. The Mahalanobis distance shows approximately linear relationship with a $\mathcal{X}_2$ distribution when the data set of concern is multinormal distributed. We analyze the distribution of T_Z and S_D in more detail later on. We can now observe that their distribution is contaminated, which, in particular, gives rise to the tail. The contamination is manifested in the Q-Q plot through divergence of the quantile equality line (the dashed line in Figure 1). We observe that the *bad* and *poor* URs correspond to the tail of the distributions, and they are separated from the main body of the distribution. Therefore the distribution can be modeled as being contaminated, where the measurement points in the tail are part of a contamination distribution and considered to be anomalies or outliers of the QoS average behavior. The contaminated part is of particular interest as this is an indicator of poor satisfaction.

C. Mahalanobis Distance Tolerance

Given that the squared Mahalanobis distance is approximately $\mathcal{X}^2_{p=2}$ distributed, points larger than $\sqrt{\mathcal{X}^2_{p,0.975}}$ can theoretically be considered as outliers [7]. To get a better insight we construct a set of vectors in $\mathbb{R}^2$ that define the $97.5\,\%$ tolerance ellipse on the T_Z-S_D plane based on the

 ISBN: 978-1-61208-126-7

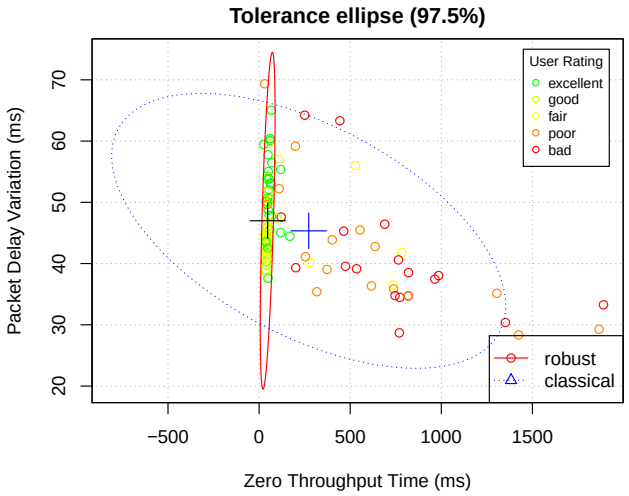


Fig. 2. The Zero Throughput Time (T_Z) - Packet Delay Variation (S_D) data plane and the 97.5 % tolerance ellipse computed with classical and robust statistical methods. The data points are colored by to their User Rating (UR).

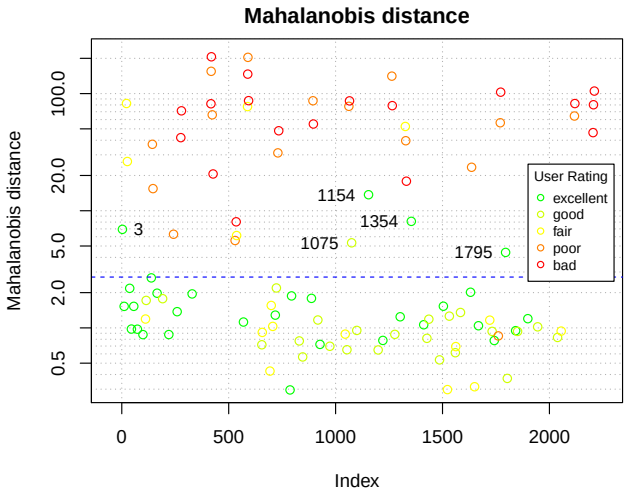


Fig. 3. The Mahalanobis distance per measurement point. The $\sqrt{\chi^2_{2,0.975}}$ cutoff is drawn with a dashed blue line. The data points are colored according to their User Rating (UR).

rated data points. This corresponds to the theoretical χ_2 distribution's 97.5 % quantile. The tolerance ellipse is shown in Figure 2. The figure shows the tolerance ellipse obtained with classical statistics and with robust statistical methods. The robust location estimation of the data set is depicted with the black crosshair, the blue crosshair is the center as specified by classical statistics. It is clear that the classical tolerance ellipse is inflated toward the outliers of, mostly, T_Z. The robust ellipse seems to cope well with the outliers. The robust estimation of the center $(47.61, 46.98)$ focuses on the major mass center of the data set, in contrast to the classical center estimation $(260.52, 45.57)$. The estimation of location for S_D is approximately the same for both methods but the T_Z is 547.19 % overestimated by the classical method compared to the robust one.

When we analyze the consistency of the URs inside the tolerance ellipse and outside we observe that inside the ellipse mostly 5 and 4 ratings are located. Outside the ellipse, the lower ratings reside. UR 3 seems to be mostly inside the tolerance ellipse. We do not observe a tendency of the URs distribution in these two areas. After interviewing the humans under investigation we observed that the streamed videos were mostly of satisfying quality, and when poor performance was noticed the satisfaction was very bad. In other words, the users perceived a reasonably good service, or a quite bad service, and rarely something in between. This is translated in a binary satisfaction, where the service perceived is either *good* (contraction of URs *excellent* and *good*) or *bad* (contraction of URs *poor* and *bad*). We can simulate such behavior by the use of the tolerance ellipse, namely inside the ellipse good service is perceived and outside bad one is perceived.

D. Spread of the Mahalanobis Distance

Figure 3 shows the Mahalanobis distance per measurement point. The points are colored according to their UR, and the horizontal dashed line is the $\sqrt{\chi^2_{p,0.975}}$ cutoff at 2.72. The *good* and *excellent*-rated measurement points above the cutoff are numbered. We observe that 91 % of the URs above 3 are under the cutoff and 97 % of the URs under 3 are captured by the $\sqrt{\chi^2_{p,0.975}}$ cutoff. As a result 9 % of good URs are located in the outlier area, i.e., points 3, 1075, 1154, 1354, and 1795. 68.75 % of the *fair* ratings are below the cutoff line. Ideally we want a clear separation between *good* and *bad* ratings. Possible causes why this is not the case is improper rating of the objects of concern or caused by the smoothing effect of the EWMA. The latter effect can be diminished by using a smaller α in equation 4. Yet, altering α also affects other measurement points. In the case of our measurements, $\alpha = 0.75$ yields optimum results. Only one value rated under 3 falls under the cutoff. Reasons for this are similar to the previous case.

E. Contamination of T_Z & S_D

The $\sqrt{\chi^2_{p,0.975}}$ cutoff divides our data set in two parts. The data points representing the network conditions yielding good user perceptions lie below the cutoff; above the cutoff are the outliners, indicators of unstable network conditions. With this information, we can model the T_Z and S_D as contaminated time series. The conventional model of contaminated data is given by

$$F = (1 - \varepsilon)\, G_0 + \varepsilon\, H, \tag{6}$$

where $\varepsilon \in [0, 0.5]$ is the degree of contamination, G_0 is the model distribution and H is the contaminating distribution. ε must be smaller than 0.5 because this is the limit where the contamination would become the model distribution and visa versa. The model in equation 6 can be applied to our QoS metrics where the data points below the cutoff are of distribution G_0 and the data point above the cutoff of distribution H. ε is defined to be the ratio of data points above the cutoff over the total number of data points, in our experiment $\varepsilon = 0.137$.

 ISBN: 978-1-61208-126-7

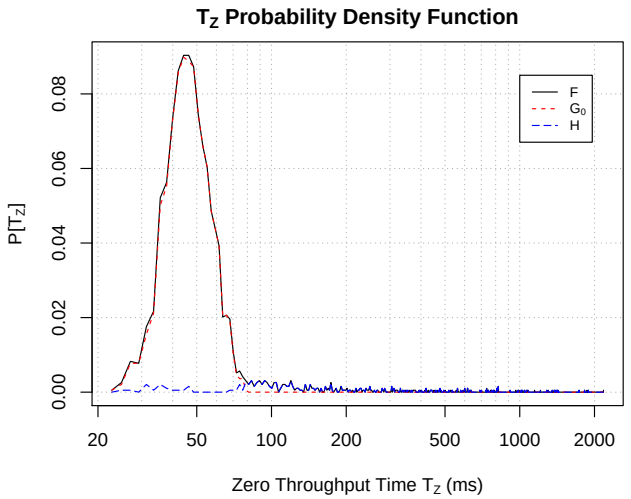


Fig. 4. Histogram of the Zero Throughput Time (T_Z), the model and the contamination distribution with a bin size of 2.15 ms.

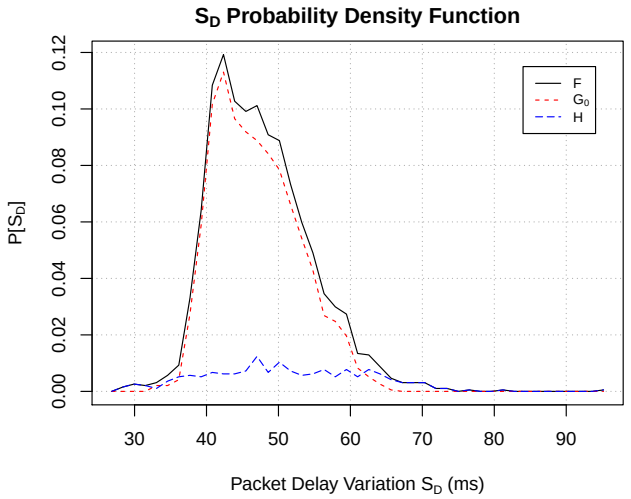


Fig. 5. Histogram of the Packet Delay Variation (S_D), the model and the contamination distribution with a bin size of 1.55 ms.

Figures 4 and 5 show the Probability Distribution Function (PDF) of T_Z and S_D, respectively. The black line is the PDF of the measured metrics, F in equation 6. The dashed red line corresponds to G_0 (the model distribution) and the blue dashed line is the identified contamination, i.e., H.

The contamination of T_Z in Figure 4 covers the tail of the F distribution. The model distribution G_0 fits well the body of F as only minor contamination is present. This suggests that the tail of T_Z can be an estimator for the QoE. The contamination H of S_D at the other hand is mixed in the whole distribution. Similar to T_Z, the contamination accounts for most of the tail and in the case of S_D also the head of F. H is considerable more present in the body of S_D than compared to T_Z. For the S_D it is impossible to identify the contamination merely on the basis of F.

When applying the $\sqrt{\chi^2_{p,0.975}}$ cutoff we assumed approximate multinormal distribution. The *Anderson-Darling* normality test on G_0 of T_Z and S_D does not yield results in favor of the normality hypothesis. Both data sets reveal positive skewness, which is larger for S_D than for T_Z. A larger data set helps to clarify the normality hypothesis. Yet, we showed that the Mahalanobis distance is an effective estimate for QoE.

V. QoE Analysis During Runtime

The Mahalanobis distance in the above analysis was computed offline after the experiments took place. Employing the Mahalanobis distance analysis during runtime implies the computation of the S matrix and the center of the data during runtime. For performance reasons this is not desirable, especially on devices with scarce resources such as handheld devices. The Mahalanobis distance computation can be simplified by using a predefined S matrix and μ vector. The computation of the Mahalanobis distance is then reduced to the multiplication of a matrix, and two vectors. The accuracy of the Mahalanobis distance estimates are consequently dependent on the selected S and μ. These values must be computed on a network with a QoS that yields good QoE most of the time. S and μ might differ for different Internet access technologies and should be studied separately before merging.

VI. Conclusions and Future Work

In this paper, we focused on the use of the Mahalanobis distance for the user satisfaction estimation of streaming video services. We showed with the help of Mahalanobis distance that there is an approximate binary relationship between QoS and QoE in 3G networks. We also showed that we can reduce the dimensions of the QoS metrics to two, without loosing much information. Of the QoS metrics measured in our experiments, T_Z and S_D are the metrics that best describe the data set with regards to its variation.

We also showed by example that robust statistical methods yield far better and reliable results than classical statistical methods. Robust statistical methods are able to handle outliers better than classical methods. In our data set, we are particularly interested in outliers, thus appropriate statistical methods are of great importance.

Future work includes the generalization of the Mahalanobis distance method in QoE modeling. We have shown that for the case of streaming over 3G networks, the Mahalanobis distance seems to be a good indicator of user satisfaction. Studies of different Internet access technologies and services will shed light on the *Mahalanobis' distance* generalized applicability.

Also, an optimized cutoff might yield better results and it is subject for future work. A larger data set than used in our paper is necessary to obtain more significant statistical results. A hysteresis approach to the binary modeling is also a good solution to prevent oscillatory behavior in binary modeling.

Acknowledgements

We would like to the thank the EU commission for funding this research under the PERIMETER STREP project.

 ISBN: 978-1-61208-126-7

REFERENCES

[1] I. Ketykó, K. De Moor, W. Joseph, L. Martens, and L. De Marez, "Performing QoE-measurements in an actual 3G network," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*. IEEE, 2010.

[2] Y. Gong, F. Yang, L. Huang, and S. Su, "Model-based approach to measuring quality of experience," in *International Conference on Emerging Network Intelligence*. Los Alamitos, CA, USA: IEEE Computer Society, 2009, pp. 29–32.

[3] O. Bradeanu, D. Munteanu, I. Rincu, and F. Geanta, "Mobile multimedia end-user quality of experience modeling," in *Digital Telecommunications, International Conference on*. Los Alamitos, CA, USA: IEEE Computer Society, 2006, p. 49.

[4] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, vol. 2, no. 1, April 1936, pp. 49–55.

[5] I.-T. R. E.800, "Terms and definitions related to quality of service and network performance including dependability," August 1994.

[6] S. Chikkerur, S. Vijay, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *to appear in IEEE Transactions on Broadcasting*, vol. 57, June 2011.

[7] R. Johnson, *Applied multivariate statistical analysis*, 4th ed., D. Wichern, Ed. Upper Saddle River, NJ [u.a.]: Prentice Hall, 1998.

[8] P. J. Rousseeuw, "Multivariate Estimation with High Breakdown Point." *Mathematical Statistics and Applications*, vol. B, pp. 283–297, 1985.

[9] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, pp. 212–223, 1998.

[10] C. Fauconnier and G. Haesbroeck, "Outliers detection with the minimum covariance determinant estimator in practice," *Statistical Methodology*, vol. 6, no. 4, pp. 363 – 379, 2009.

[11] S. Ickin, K. De Vogeleer, M. Fiedler, and D. Erman, "On the choice of performance metrics for user-centric seamless communication," in *Third Euro-NF IA.7.5 Workshop on Socio-Economic Issues of Networks of the Future*, vol. 3, Gent, Belgium, December 2010, pp. 4–5.

[12] F. Guyard and S. Beker, "Towards real-time anomalies monitoring for qoe indicators. annals of telecommunications," in *Annals of Telecommunications*, vol. 65. No. 1–2, Jan/Feb 2010, pp. 59–71.

[13] Apple, "http://developer.apple.com/opensource/server/streaming," Open Source Streaming Server, 2011.

[14] H. Schulzrinne, A. Rao, and R. Lanphier, "Real time streaming protocol (*RTSP*)," RFC 2326, Internet Engineering Task Force, April 1998. [Online]. Available: http://www.ietf.org/rfc/rfc2326.txt

[15] S. Ickin, K. De Vogeleer, M. Fiedler, and D. Erman, "The effects of packet delay variation on the perceptual quality of video," in *4th IEEE Workshop On User MObility and VEhicular Networks (On-MOVE 2010)*, Denver, Colorado, USA, October 2010, pp. 679–684.

[16] M. Hubert, P. Rousseeuw, and Vanden, "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, vol. 47, pp. 64–79, 2005.

 ISBN: 978-1-61208-126-7

Implementation of a Media Aware Network Element for Content Aware Networks

Dragoş S. Niculescu, Mihai Stanciu, Marius Vochin, Eugen Borcoci
Telecommunication Dept., ETTI,
University Politehnica of Bucharest
Bucharest, Romania
e-mail:{dniculescu,ms,mvochin,eugenbo}@elcom.pub.ro

Nikolaos Zotos
Institute of Informatics and Telecommunications
NCSR Demokritos
Athens, Greece
e-mail: nzotos@iit.demokritos.gr

***Abstract*—ALICANTE is a recently proposed architecture that enables a lightweight form of virtualization for the purpose of offering QoS to media streams across the internet. A content aware network (CAN) is a cross domain overlay, which is provisioned in advance in order to provide preferential treatment to media streams. While it uses legacy infrastructure, such as core IP/MPLS routers and provisioned links, it relies on a special border router, called MANE (Media Aware Network Element). This paper presents a modular implementation of such a network element using off-the-shelf hardware and open source software. The implementation uses Click modular router to implement flow classification, MPLS encapsulation and decapsulation, separation between virtual CANs, and enforcement of separation between networks. Based on incipient measurements in a virtual testbed, we show that the implementation imposes minor overheads over existing routing infrastructure.**

Keywords—Content-Aware Networking; Network Aware Applications; Quality of Services; Multimedia distribution, Future Internet; Media Aware Network Element

I. INTRODUCTION

One of the new paradigms of the Future Internet (FI) is "content orientation", which is supposed to improve the user experience related to the new digital multimedia services and networked media content. This trend is recognized also by the European commission, which defined the "Objective ICT-2009.1.5: Networked Media and 3D Internet" in the FP7 Call 4 [1][2]. In this call, new directions are defined as content-aware networks (CAN) and network-aware applications (NAA). This approach breaks (partially) the classic TCP/IP and OSI stack network neutrality and application-transport separation concepts. The challenge is to get better performance without losing modularity of the architecture. CAN-NAA means the capability of the overall system to adjust network resource allocation based on limited examination of the nature of the content, while network-awareness means to process and distribute the content, based on limited understanding of the network conditions. Dynamic optimization is desired, with policies taking into account the content and adaptation needs, the user contexts, requirements and social relational networks. The FI should enable multiple user roles, e.g., as content producer, user, or manager.

The work of this paper is a part of the effort inside of an European FP7 ICT research project, "MediA Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments", ALICANTE, [3][4]. An innovative architecture is proposed, in order to deploy a new type of "Networked Media Ecosystem", which allows flexible cooperation between providers, operators, and users. Three interworking environments are defined: Network Environment (NE) modeling Network Providers, Service Environment (SE) related to Content and Service Providers and User Environment (UE) including all End-Users (see Figure 1). The validation of the project architecture and results will be done in a large-scale international pilot.

The above environments are nowadays present in real deployments, but actually the collaboration between them is weak or non-existent. The current architectures do not exchange content-based and network-based information between the network layers and upper layers. This neutral network was considered many years as a strong principle governing the Internet, however it begins to show some disadvantages taking into account the multimedia-intensive aspects of the FI.

The paper is organized as follows: Section II presents some of the related work existent in the field. The ALICANTE architecture and its main concepts are defined in Section III. Section IV is focused on the MANE implementation, while some measurements are presented in Section V. Conclusions, open issues, and future work are presented in Section VI.

II. RELATED WORK

The content-aware networking (CAN) and network-aware applications (NAA) approach is a new mode to design the layered architecture, with a running debate about the benefits of better interactions as opposed to the penalty of losing modularity of the architecture.

 ISBN: 978-1-61208-126-7

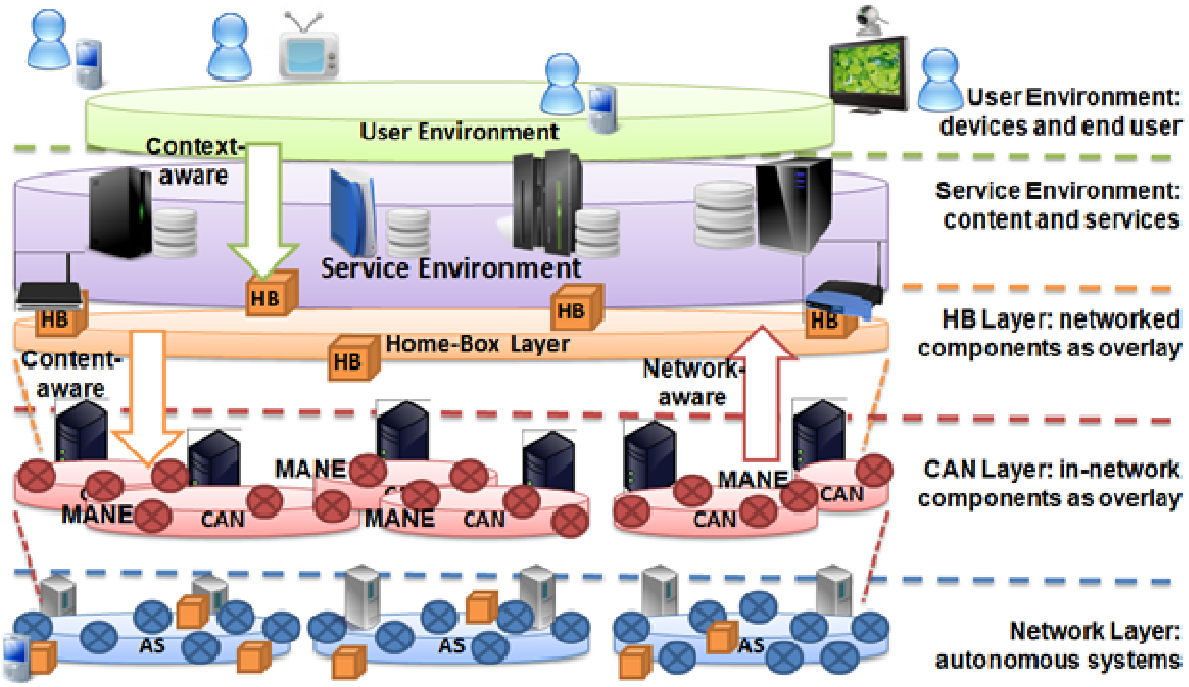


Figure 1. The ALICANTE Architecture

In such a context, both CAN and NAA are of interest both for research communities and industry, in the process of re-thinking the architecture of the FI.

The capability of content-adaptive network awareness to offer optimization for video transmission is analyzed in [5]. In [6], it is considered that CAN and NAA can offer a way for evolution of networks beyond IP. In [7], it is discussed how the CAN/NAA approach can lead to a user-centric FI and telecommunication services. The content adaptation issues in the FI as a component of CAN/NAA approach is discussed in [8]. The better QoE/QoS capabilities of the CAN/NAA architecture is analyzed in [9][10]. Further gains are obtained if context awareness is also considered [11][12].

Conversely, packet header processing time in the CAN routers raises concerns similar to the deep packet inspection techniques problems [13]. The application layer traffic optimization (ALTO) problem defined by the IETF can be solved by the cooperation between the CAN layer and the upper layer, as in [14][15].

However, no complete and open architecture currently exists, able to support multimedia distribution according to the CAN principles and scalable over sizeable networks and heterogeneous networking technologies. Therefore, an open field for research in this domain exists.

III. ALICANTE System Architecture

A. *Layers and entities*

The ALICANTE architecture, as shown in Figure 1, promotes concepts such as content-awareness to the network environment, user context-awareness to the service environment, and adapted services/content to the end-user for his/her best service experience while being either a consumer and/or producer.

Two new virtual layers are proposed on top of the traditional network layer: the CAN layer for network level packet processing and a Home-Box (HB) layer for the actual content delivery.

The CAN layer offers an enhanced support for packet payload inspection, processing and caching in network equipment. It is developed over traditional IP network/transport layer. It will improve data delivery via classifying and controlling messages in terms of content, application and individual subscribers; it improves QoS assurance via content-based routing and increases network security level via content-based monitoring and filtering. In such a way, content- and application-aware networks are created to provide high levels of performance, end-user experience, and to enable application and subscriber-specific data forwarding. The specific components in charge of creating this CAN layers are the Media-Aware Network Elements (MANE), i.e., the new CAN routers, and the CAN managers.

The Home-Box layer is an upper layer, using CAN services and taking into account network-aware information delivered upward by the CAN layer. Thanks to this layer, inter-working with the User, Service, and Network Environments, one can elaborate network and context-aware applications and deliver the necessary inputs to create content-aware networks. The Home-Box (HB) is a physical and logical entity located at end-user's premises. The adaptation, service mobility, security, and overall management of services and content are being assured at this layer through a new specific middleware

 ISBN: 978-1-61208-126-7

proposed by the project, working in conjunction with the other layers.

The interactions between the above mentioned two layers establish together a powerful cross-layer optimization loop providing end-users with the best possible service experience and optimizing the resource usage.

The upper SE layer uses information delivered by the CAN layer and enforces network-aware applications procedures necessary to perform the adaptation of the media resources to the user's preferences.

The main management and control entity in the CAN layer is the CAN Manager (CANMng). Corresponding to its roles, we distinguish the following interfaces of CAN Manager with the Virtual Home-Box layer: to advertise CANs and negotiate their usage and to help the establishing of connectivity relationships at Virtual HB layer based on, e.g., network related distance information. The CANMng has also interfaces to the lower network layer in order to negotiate CANs and request their installation.

Each AS has one CANMng, playing the following roles: to (re)define the CANs (according to the enhanced connectivity service targeted) and perform all related actions to configure, maintain and update CANs; to advertise and negotiate the CAN usage with upper layers, using Service Level Agreements/Specifications (SLA/SLS) contracts; to communicate with other CAN managers in order to establish multi-domain chains, again, using SLA/SLS contracts; to communicate with its own intra-domain network resource managers (IntraNRM). The IntraNRMs have the ultimate authority upon the network provider resources, thus conserving each domain's independency.

B. The Content-Aware Network Router

The MANE, a content-aware network router, is an intelligent network node. It performs appropriate processing (routing, filtering, adaptation, security operations, etc.) taking into account the content type, the content properties (described by metadata or extracted by protocol field analysis) and also depending on network properties and network status. The results of the content related information analysis provide metrics, which help deciding the best strategy to adopt for the best content repurposing and publishing methods. The MANE basic set of functions are:

Content-aware intelligent routing: the MANE will decouple the higher level routing process from the lower level forwarding and perform intelligent routing, based on results extracted from packet fields analysis or content description metadata

Content-aware QoS and resource allocation: the MANEs will be able to deduct the QoS requirements of different flows based on the flows content. The CAN layer will load-wise monitor the current status of the CANs. The MANE will maintain an aggregated image of flows that they forward, and for every recognized flow type, an instance of CAN (VCAN – Virtual CAN) will be assigned depending on the level of QoS guarantees and network status. This will optimize resource allocation in the network depending on traffic types and QoS requirements The CAN level will interact with the domain network resource management in order to perform mapping onto different L2/L3 QoS-aware technologies (e.g., MPLS/Diffserv or Carrier Ethernet). Some amount of relatively infrequent dynamic re-allocation of the network resources between different CANs is possible, optimizing resource usage. The MANE has also an adaptation role, deployed at different points in the delivery chain: at the service creation, during the transport by the CAN routers, and at the Home-Box site;

Specific Security issues: increase usage of encryption technologies (such as IPSec) at network level has the direct consequence that content type can become hidden and packet inspection becomes ineffective; to mitigate this, we will exploit the possibility to include content related information in dedicated fields. Thus, the end-to-end communication remains encrypted and private, while the content-aware network concept can still function.

Another issue, beyond privacy, which is addressed by using special fields and/or metadata to describe the content, is the processing time required by deep packet inspection; eliminating the need for this procedure will significantly improve the performance of the CAN-enabled routers.

IV. MANE HIGH LEVEL ARCHITECTURE

The main interactions of the MANE are presented in Figure 2. The control plane interactions are indicated with thick shaded arrows, and the data path is indicated with thin arrows. From above, the Intradomain NRM has the role of providing the means to create FEC associations for entry in the MPLS domain. Each packet is marked with a VCAN header by the generating HB/SB, and a path is decided through CAN Manager – IntraNRM collaboration. The type of VCAN and the entry into the MPLS tunnel are then provided for each entry MANE router. The core of the mane is a classifier/router module which identifies incoming traffic based on its VCAN header, encapsulates it into the MPLS header, and sends it to the appropriate LSP.

 ISBN: 978-1-61208-126-7

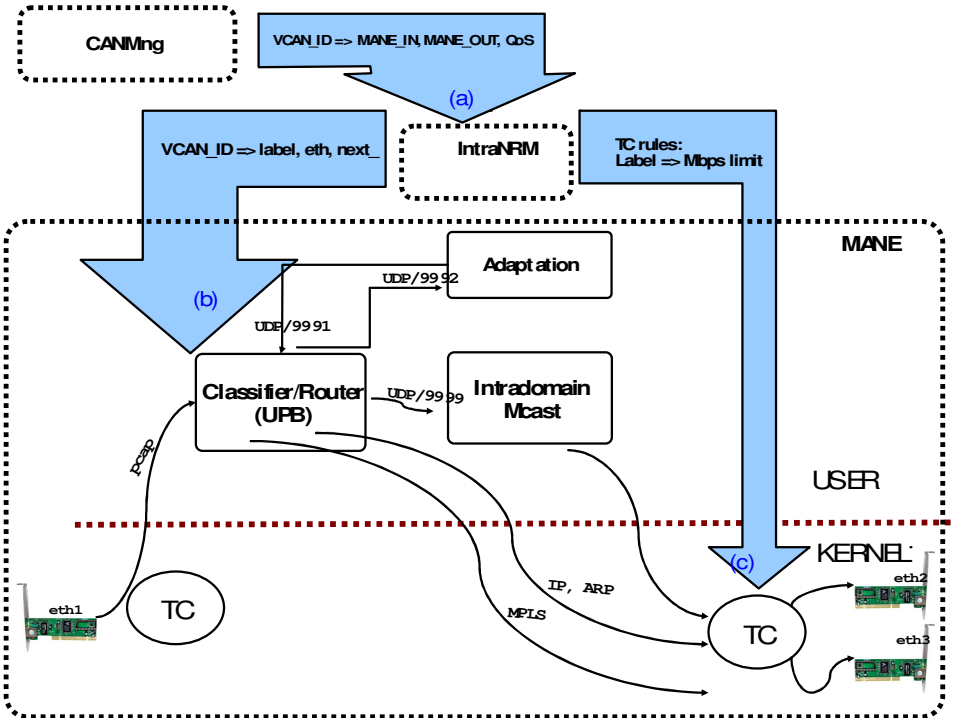


Figure 2. The ALICANTE Architecture: details on Virtual CAN Layer

The architecture is completely modular so that functionality can be developed in parallel. Modules all run in user space and are interconnected using UDP/IP so they can run on either the same, or on different machines. Another advantage is that it is easy to reconfigure the architecture by simply changing the UDP ports, or by inserting new modules.

The central module is the Classifier/Router, which harvests packets from the incoming interface and distributes them to the other modules, or encapsulates them into the MPLS paths. The classifier needs to know the association between VCAN IDs present in all incoming packets so that it can route traffic to the appropriate VCAN. VCANs are assumed to be configured in advance by the CAN Manager and provisioned through the IntraNRM. In particular, MANE needs to be instructed explicitly by the IntraNRM on the association between the VCAN_ID and an MPLS label to be used (shaded arrow marked b)). This module also decapsulates MPLS traffic that comes from the domain, before forwarding it to the appropriate HB/SB.

The core router part is not represented here, but we assume it has complete MPLS support and is implemented either with specialized hardware, or with Linux machines requiring a specially patched kernel. The IntraNRM manages al the labels and the bandwidth provisioning for each path. In fact bandwidth provisions are sent down to both MANEs and core routers to be enforced, perhaps with `tc` functionality (marked as shaded arrow c).

The Classifier/Router module is implemented in user space and uses Click modular router [16], as shown in Figure 3. For close to Linux performance it could be moved down into the kernel space in the final phases of development. It performs MPLS and IP routing both in and out the domain. It performs FEC associations for incoming IP traffic, and MPLS decapsulation for traffic outgoing to HB. It dispatches traffic to local modules (Adaptation, multicast, etc), but also accepts traffic from them so they don't need to handle routing or encapsulation tasks. The convention in implementing the MANE is that eth0 interface is used for testbed support and therefore not part of ALICANTE. Interfaces eth1, eth2, eth3 … are used either as ingress into, or egress from the MANE. The main classification task is performed by a dedicated classifier element, called `cl_ing` (for ingress traffic), that identifies the VCAN of the incoming packet and uses the appropriate MPLS label to decide the policies for forwarding, shaping and policing. The elements grouped in the `elementclass` Card handle all the bookkeeping necessary to IP and MPLS to exchange packets on the local network (for card2 and card3, the internal details are omitted).

 ISBN: 978-1-61208-126-7

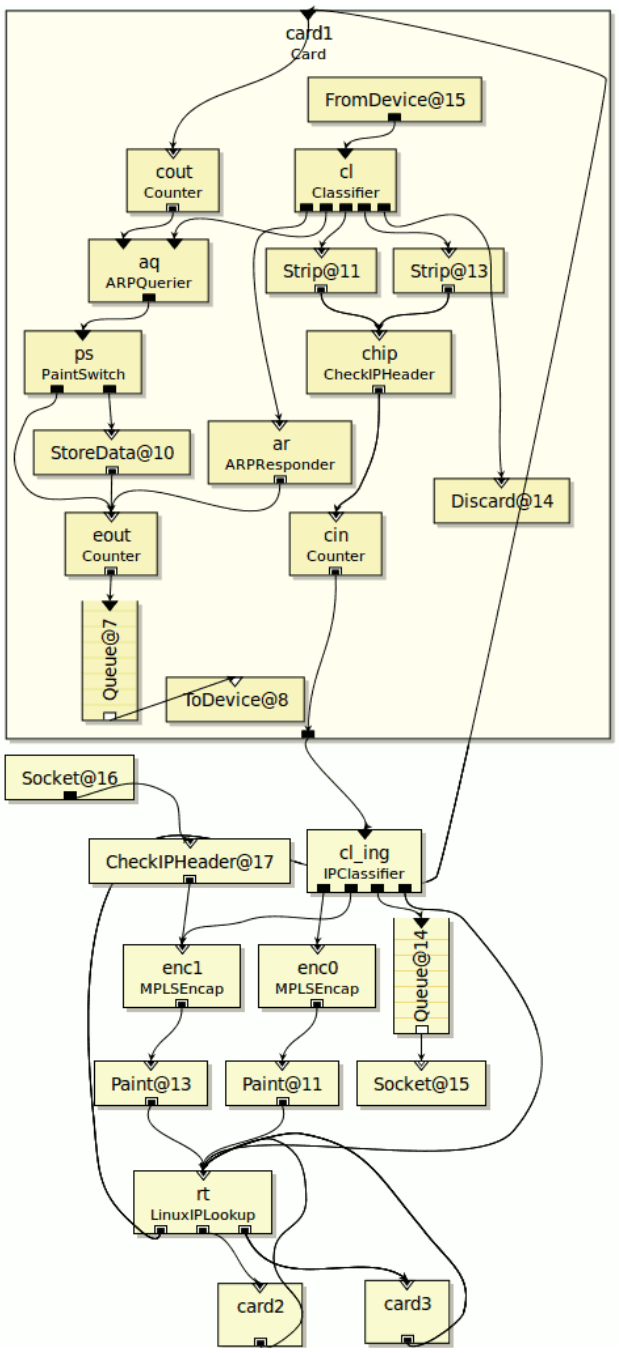


Figure 3. Userspace MANE classifier/router implementation using Click elements

V. EXPERIMENTS

Using the implementation of MANE described in the previous section, we built a topology comprising three HBs, three MANEs, and two MPLS core routers, as shown in Figure 4.

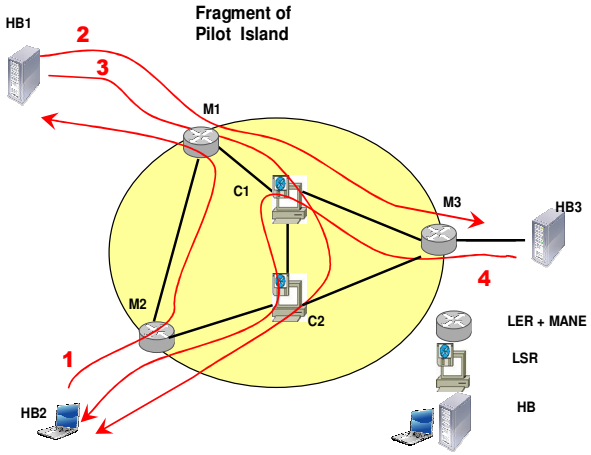


Figure 4. Pilot island using three MANE routers, two core routers and three home boxes (HB).

At this phase of the project, all the elements in the picture run in a virtualized fashion using VirtualBox on a server with quad CPU Intel Xeon X2250, 4GB memory, running Ubuntu Linux. We measured performance for three different routing configurations: using standard IP, using kernel MPLS, and using our implementation of MANE (currently in user space). The results for path 3 are summarized in Table I:

TABLE I. MEASUREMENT RESULTS

	IP	MPLS	MANE
ping 32 byte packets RTT/stddev [ms]	1.46/0.44	1.48/0.45	1.70/0.69
ping 1460 byte packets RTT/stddev [ms]	1.53/0.46	1.43/0.65	1.75/0.78
UDP Rate[Mbps]/Load	45.3/66%	45.2/66%	41.0/85%
TCP Rate[Mbps]/Load	61.0/33%	60.0/37%	37.8/62%

For the RTT tests, we used ping with large and small packets. The MANE implementation brings a minor increase in end to end transit time and a slight increase in the standard deviation of RTT. The data rates achieved with the MANE implementation are 10% less than the plain Linux data rates, but part of this difference can be accounted by the current implementation of all modules in user space. This also explains the increase in processing time in the networking elements, reported as 'Load' in the table.

VI. CONCLUSIONS AND FUTURE WORK

We presented the high level architecture of a media aware network, which aims at virtualizing network resources for the purpose of offering higher QoS to media flows. The MANE (Media Aware Network Element) is an

 ISBN: 978-1-61208-126-7

edge router that has a central role in implementing the separation between networks, by classifying incoming traffic and distributing it to appropriate MPLS paths inside each domain.

We implemented the MANE using off the shelf hardware, using Click modular router to interface the components: classification, routing, adaptation, multicast, MPLS FEC association, encapsulation and decapsulation. Our preliminary implementation on a virtual testbed shows a modest increase in processing overhead when compared with traditional IP/MPLS processing.

In the future we aim at developing the MANE in two directions: adding deep packet inspection functionality, to assist in classification of traffic not yet associated with a VCAN, and integration with high speed network processing cards, to target operation at line speed. Both these directions aim at creating a MANE that can be deployed in the field by service providers.

ACKNOWLEDGMENTS

This work was supported in part by project POSDRU/89/1.5/S/62557 and in part by the EC in the context of the ALICANTE project (FP7-ICT-248652)

REFERENCES

[1] Networked European Software and Sevices Initiative (NESSI) Strategic Research Agenda, Vol. 3. FP7-2.exec, NESSI Roadmap, Feb. 2008.

[2] European Commission, FP7 ICT Work Programme 2009-2010.

[3] Borcoci, E., Negru, D. and Timmerer, C., "A Novel Architecture for Multimedia Distribution based on Content-Aware Networking" Proc. of. CTRQ 2010, Athens, June 2010, pp. 162-168.

[4] FP7 ICT project, "MediA Ecosystem Deployment Through Ubiquitous Content-Aware Network Environments", ALICANTE, No248652, http://www.ict-alicante.eu (last accessed: Mar. 2010).

[5] Martini, M.G., Mazzotti, M., Lamy-Bergot, C., Huusko, J., and Amon, P., "Content Adaptive Network Aware Joint Optimization of Wireless Video Transmission", IEEE Communications Magazine, vol. 45, no. 1, Jan. 2007, pp. 84-90.

[6] Kourlas, T., "The Evolution of Networks beyond IP", IEC Newsletter, vol. 1, Mar. 2007. Available at www.iec.org/newsletter/march07_1/broadband_1.html (last accessed: Mar. 2010).

[7] Baladrón, C., "User-Centric Future Internet and Telecommunication Services", in: G. Tselentis, et. al. (eds.), Towards the Future Internet, IOS Press, 2009, pp. 217-226.

[8] Zahariadis, T., Lamy-Bergot, C., Schierl, T., Grüneberg, K., Celetto, L., and Timmerer, C., "Content Adaptation Issues in the Future Internet", in: G. Tselentis, et al. (eds.), Towards the Future Internet, IOS Press, 2009, pp.283-292.

[9] Liberal, F., Fajardo, J.O., and Koumaras, H., "QoE and *-awareness in the Future Internet", in: G. Tselentis, et al. (eds.), Towards the Future Internet, IOS Press, 2009, pp. 293-302.

[10] Huszák, Á. and Imre, S., "Content-aware Interface Selection Method for Multi-Path Video Streaming in Best-effort Networks", Proc. of 16th International Conference on Telecommunications, Marrakech, Morocco, Jul. 2009, pp. 196-201.

[11] Baker, N., "Context-Aware Systems and Implications for Future Internet", in: G. Tselentis et. al. (eds.), Towards the Future Internet, IOS Press, 2009, pp. 335-344.

[12] Kodeswaran, S. B. and Joshi A., "Content and Context Aware Networking Using Semantic Tagging", Proc. of 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, Georgia, USA, Apr. 2006, pp. 67–77.

[13] Rainge, E., "The Inevitable Failure of Content-Aware/DPI Network Devices and How to Mitigate the Risk", Sept. 2008 (adapted from Worldwide Network Test and Measurement 2008.2012 Forecast and 2007 Market Shares, Available at http://www.breakingpointsystems.com/resources/white-papers/idc-white-paper/content-aware-testing.pdf (last accessed: Mar. 2010).

[14] Aggarwal, V. and Feldmann, A., "Can ISPs and P2P Users Cooperate for Improved Performance?", ACM SIGCOMM Computer Communication Review, vol. 37, no. 3, Jul. 2007, pp. 29-40.

[15] Xie, H., Krishnamurthy, A., Silberschatz, A., and Yang, Y., "P4P: Explicit Communications for Cooperative Control Between P2P and Network Providers", Available at http://www.dcia.info/documents/P4P_Overview.pdf (last accessed: mar. 2010)

[16] Kohler E., Morris, R.T., Chen B., Jannotti J., and Kaashoek, M. F., "The click modular router", ACM Trans. Comput. Syst. 18, 3 (August 2000), pp. 263-297

 ISBN: 978-1-61208-126-7

FTAM: A Fuzzy Traffic Adaptation Model for Wireless Mesh Networks

Ali El Masri, Lyes Khoukhi, and Dominique Gaiti
ICD/ERA, Troyes University of Technology
12 rue Marie Curie, 10010 Troyes Cedex , France
Email: {ali.el_masri, lyes.khoukhi, dominique.gaiti}@utt.fr

***Abstract*—Wireless mesh networks (WMNs) are considered as the next step towards providing a high-bandwidth network over a specific coverage area. Because of their advantages over other wireless networks, WMNs are undergoing rapid progress and inspiring numerous multimedia applications such as video and audio real-time applications. These applications usually require time-bounded service and bandwidth guarantee. Therefore, there is a vital need to provide Quality of Service (QoS) support in order to assure better quality delivery. However, providing QoS support for real-time traffic in WMNs presents a number of significant technical challenges. In this paper, we focus on one of the most critical technical issues in QoS support, by proposing a novel QoS traffic adaptation model based on fuzzy logic theory, named FTAM, which is capable of supporting real-time traffic such as video and voice services. By monitoring the rate of change in queue length in addition to the current length of the queue, FTAM is able to provide a good measurement of the future queue state, and then to achieve the convenient traffic adaptation according to the network state.**

***Index Terms*—Wireless Mesh Networks; QoS; Traffic Adptation; Fuzzy Logic;**

I. Introduction

The last few years have witnessed a wealth of research ideas on Wireless Mesh Networks (WMNs) that are moving rapidly toward implemented standards. Although WMN research is a relatively new field it is gaining more popularity for various new applications. For instance, multimedia application that opens up for converged services and new purposes is quickly becoming a key focus area for wireless mesh communications [1]. With the increase in both the bandwidth of wireless channels and the computing power of mobile devices, it is expected that video and audio services will be offered over WMN in the future. However, enabling multimedia communications over such networks is remaining a challenging task for both academic and industrial communities. Video and audio real-time services typically require stringent bandwidth and delay guarantees. This makes the deployment of Quality of Service (QoS) mechanisms a vital need for the satisfaction of user's requirements. Real-time applications generate traffic at varying rates and usually require the network to be able to support such a changing rate. Therefore, providing QoS guarantees is crucial for supporting disparate services envisioned for future wireless mesh networks [2].

Despite the efforts made to alleviate this issue, there still exist a number of barriers to the widespread deployment of real-time applications. The most prominent one is how to ensure the traffic adaptation in the case of heavy congestion case. It is important to note that the existing solutions developed for wired networks can not be deployed directly within WMNs. Difficulties with these models lie in the fact that they are not adapted to different node states and resource variation, as in mesh environments the available bandwidth for each node varies with time since the medium is shared [3].

In this paper, we introduce a novel QoS model for traffic adaptation based on fuzzy logic that is capable of supporting real-time traffic such as video and voice services. A major factor behind using fuzzy logic theory to ensure the traffic adaptation, is its adequation to the uncertainty, the heterogeneity and the information incompleteness of WMN environment characterized by dynamic traffic changes. Our proposed model, build on both MAC and network layers, will bring about the benefits of the advances in the areas of artificial intelligence and wireless networking. The evaluation of the model performances will be studied under different traffic and network conditions. The balance between network performances and reliability when transmitting multimedia traffic is an important issue to consider too.

The remaining of the paper is organized as follows. A brief description of Fuzzy Logic theory is provided in Section II. Section III gives a state of the art regarding QoS fuzzy models in WMN. Section IV presents our proposed model. In Section V, we discuss the performance evaluation of FTAM, while Section VI concludes the paper.

II. Fuzzy Logic

In this section, we give a brief overview of the Fuzzy Logic theory to help the unfamiliar reader to understand the rest of the paper. Exhaustive description can be founded in the literature [4] [5].

A. Fuzzy Sets

Fuzzy sets represent a modernization of traditional crisp sets where the membership of an object x in a set A is evaluated by 1 (true) or 0 (false). True signifies that x is member of A and false signifies that x is non-member of A. Fuzzy sets allow the partial membership of x in A. The degree of membership has a real value in [0, 1], where 0 and 1 correspond respectively to the full non-membership and the full membership of x in A. If A is a fuzzy set in a universe U, the membership of x in A is evaluated by the membership function μ_A as following:

 ISBN: 978-1-61208-126-7

$$\mu_A : U \rightarrow [0, 1]. \quad (1)$$

Each $u \in U$ has a degree of membership in A equal to $\mu_A(u)$. An object x is defined as a linguistic variable such as distance or speed, and a fuzzy set A is defined as a linguistic term such as far or high.

B. Fuzzy Logic Controllers

A *Fuzzy Logic Controller* (FLC) is a tool used to compute the value of an output based on several inputs having between them a complex relation which cannot be solved via traditional mathematical tools such as weighted sum. The structure of a FLC is shown in Figure 1. To compute the value of the output of a FLC, the following steps must be applied.

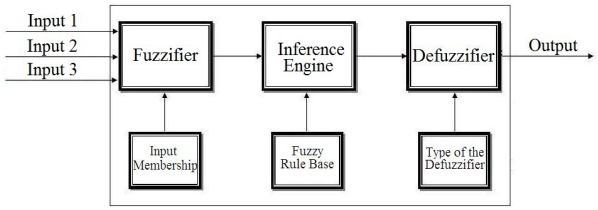


Fig. 1: Fuzzy Logic Controller

First, we compute or measure from the external environment the value of each input. Second, these values are converted by the *fuzzifier* into fuzzy variables ready to be used by the *inference engine*. This step is called *fuzzification* and it is executed based on the *membership functions* of each input. An example of *membership functions* is shown in 1. Third, the *inference engine* applies each rule of the *fuzzy rule base* to the input fuzzy variables to compute an output fuzzy variable. An example of *fuzzy rule base* is shown in Table I. The rules of the *fuzzy rule base* have the following form: IF (*input 1* is X_1 and *input 2* is X_2 and *input 3* is X_3) THEN (*output* is Y). Fourth, the output fuzzy variables of all the rules are connected to compute the final output fuzzy variable. Finally, the final output fuzzy variable is converted by the *defuzzifier* into a crisp output ready to be used in the external environment.

III. Fuzzy Logic In Wireless Multihop Networks

Fuzzy logic has been successfully applied to resolve problems that are either difficult to tackle mathematically or where the use of fuzzy theory provides improved performances [6]. In what follows, we present some relevant QoS models proposed in the literature.

In [7], we proposed an integrated stateless cross-layer QoS protocol FuzzyQoS based on fuzzy logic for wireless mobile ad hoc networks. The choice of using fuzzy logic is justified by the fact that fuzzy logic is well adapted to systems characterized by imprecise states, as in the case of ad hoc networks. The fuzzy approach aims to improve the control of traffic regulation rate and congestion control of multimedia applications. FuzzyQoS uses fuzzy thresholds to adapt the traffic transmission rate to dynamic conditions. The performance evaluation has shown that FuzzyQoS can achieve low and stable end-to-end delay, and high throughput under different network conditions.

In [8] a fuzzy logic based cooperative MAC protocol (FLCMAC) is proposed to cooperate amongst network flows and dynamically adjust access probability of each low priority flow affecting the high priority flows to satisfy their QoS requirement. The simulation results have indicated that compared to the enhanced distributed channel access (EDCA) scheme of 802. 11e, the FLCMAC gives better performances in terms of throughput and delay under moderate and heavy background traffic both in single-hop and multi-hop scenarios. This work addresses the problem of spatial reference estimation in mobile scenarios.

In [9], the authors proposed a new model to investigate the use of fuzzy logic theory for assisting the TCP error detection mechanism in an ad hoc network. An elementary fuzzy logic engine was presented as an intelligent technique for discriminating packet loss due to congestion from packet loss by wireless induced errors. The results have shown that the fuzzy engine may distinguish congestion from channel error conditions, and consequently assist the TCP error detection [9]. Reznik *et al.* in [10] have investigated the issues for improving the reliability and accuracy of the decisions in wireless ad hoc networks. They proposed an approach that offers a way of integrating wireless units measurement results with association information available or priori derived at aggregating nodes. This approach is used for describing both wireless units results and association information with consideration given to both Neuro-Fuzzy and probabilistic models and methods. The information sources available in the system are classified according to the model (fuzzy or probabilistic), which seems more feasible to be applied [10].

IV. Fuzzy Traffic Adaptation Model

An efficient network congestion control has to prevent the packets losses, which are caused by unexpected traffic bursts. Thus, it has to estimate the dynamic behavior of the traffic in the nodes buffers and to send sources the congestion notifications early enough. Therefore, due to the dynamic nature of buffer occupancy and congestion at a node, we expect that applying a fuzzy logic control seems to be a very interesting issue.

In this section, we propose a *fuzzy logic controller* FLC-FTAM (Figure 2), which is designed to offer a better adaptability under varying network conditions by better tuning of rules without intervention of operators.

The proposed model is conceived as a nonlinear controller in which the input-output relationship can be expressed by using a small number of linguistic rules or relational expressions. The goal of our proposal is to make rate control decisions based on the instantaneous queue length and the variation rate of the queue length at each wireless node. By monitoring the rate of changes in queue length (variation rate) in addition to the queue length, the model is able to provide a measure of queue state, and by using explicit rate congestion notifications

 ISBN: 978-1-61208-126-7

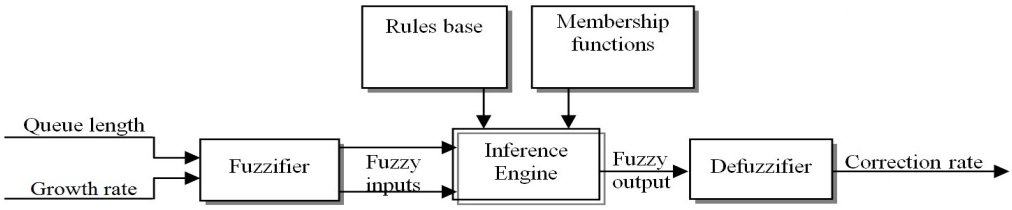


Fig. 2: FTAM model

we can make source nodes more responsive to sudden changes in the network traffic volume.

Let's consider that the current queue length of a node is around the half of the queue size and the rate changing of the queue length is during an increasing phase, then the queue will be filled in the near future. Hence, a congestion may occur and future arrival packets may be dropped. To prevent congestion, the incoming flow rate should be immediately decreased. In the other case, when the current queue length of a node is close to zero and the rate changing of the queue length is during a decreasing phase, then the flow rate should be increased to fully optimize the utilization of radio resources and to fully maximize the overall throughput of the wireless network.

To illustrate well this concept, we use the following rule (all rules used by our model to achieve the traffic adaptation process are shown in Table I):

"If the rate changing in the buffer is rising fast and the buffer is filled, then the flow rate should be very-little".

In this expression, *"buffer"*, *"rate changing"* and *"flow rate"* are called linguistic variables which accept values among the words of a natural or synthetic language such as *"filled"*, *"rising fast"* or *"very-little"*. Usually, the values related to the linguistic variables are modeled by means of fuzzy sets. In most cases, this way of representation offers both a good description of systems and a natural behaviour of any inherent non linearity in the control process. The design of FTAM involves the selection of suitable mathematical representation of fuzzification and defuzzification operators, fuzzy implication functions, and forms of membership functions among a set of candidates. Particular choice of these functions and operators may affect the behavior of the traffic adaptation controller.

Figure 2 illustrates the structure of the proposed *fuzzy logic controller* FLC-FTAM. FLC-FTAM uses two input parameters to compute an explicit traffic rate: queue length (QL) and its growth rate (GR). The value of GR is computed as the difference between the current queue length and the queue length from the previous control interval (i. e. queue growth rate). Based on the values of GR and QL parameters, and the information stored in the traffic rules base, FTAM can calculate the required change in the session rate and store this information in a field named *Explicit Rate "ER"* in a newly created congestion notification packet (*CNP*). Within the present control interval, the congested node sends a *CNP* packet that will travel to the upstream nodes along the route.

In order to obtain the *flow correction rate*, we define the fuzzy Rule Base shown in Table I. This table is a proposal for the FLC-FTMA determined via the analysis in the previous section but also by observations during simulations. Note that the rule base is malleable enough so that other researchers can argue and propose different rules for different reasons.

TABLE I: FTAM Fuzzy Rule Base

	IF		THEN
n	Growth Rate	Queue Length	Flow Rate
1	*Negative*	*Very Small*	*Increase sharply*
2	*Negative*	*Small*	*Increase sharply*
3	*Negative*	*Medium*	*Increase moderately*
4	*Negative*	*Big*	*Do not change*
5	*Negative*	*Very Big*	*Do not change*
6	*Acceptable*	*Very Small*	*Increase sharply*
7	*Acceptable*	*Small*	*Increase moderately*
8	*Acceptable*	*Medium*	*Do not change*
9	*Acceptable*	*Big*	*Decrease moderately*
10	*Acceptable*	*Very Big*	*Decrease sharply*
11	*Positive*	*Very Small*	*Do not change*
12	*Positive*	*Small*	*Do not change*
13	*Positive*	*Medium*	*Decrease moderately*
14	*Positive*	*Big*	*Decrease sharply*
15	*Positive*	*Very Big*	*Decrease sharply*

Note that in the implementation phase of FTAM model, the choice of traffic rules is performed depending on the manner how the system should behave to ensure the traffic adaptation process. After classification of the appropriate rules, the membership functions associated to each parameter's rule are identified. For that aim, a variety of membership functions may be applied to ensure the adaptation process such as triangular, Gaussian, and trapezoidal functions. In FTAM, we have chosen triangular and trapezoidal functions because of their simplicity in computation. After that, the rule base is fine tuned by observing the progress of simulation in order to achieve a suitable balance between a tolerable average end-to-end delay and increase the throughput.

V. Performance Evaluation

A. FTAM Analysis

Before simulating our proposal by computer-based simulation, we should first validate the effectiveness of using the fuzzy logic within the proposed traffic adaptation model. Since our model is based on IF-THEN rules rather than on mathematical equations, we need to show how the traffic adaptation changes as function of fuzzy inputs.

Figure 3 shows the relation function between the output (*flow correction rate*) of the model and its two inputs (*queue length* and *growth rate*). We observe that when the *queue* is

 ISBN: 978-1-61208-126-7

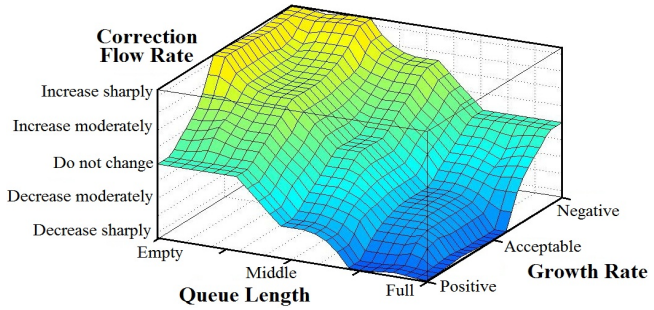


Fig. 3: Correction Flow Rate as a function of Queue Length and Growth Rate

full or empty, the *flow correction rate* is less dependent on the *growth rate*. For instance, if the *queue* is full, then the *flow rate* should be decreased except if the *growth rate* is negative. However, when the *queue* is empty, the *flow rate* should be increased except if the *growth rate* is positive. In the other case, we observe that when the *queue length* is middle, the *flow rate* is more dependent on the *growth rate*. If the *growth rate* is positive, acceptable or negative, then the *flow rate* should be decreased, not changed or increased, respectively.

The previous results typically express the relation between the *queue length* and the *growth rate*. As explained above, the target *queue length* in our model is equal to the *Medium* value of *queue length* (here equal the half of the buffer size). For instance, when the *queue* is empty, the *flow rate* should be increased until the half of the buffer size will be filled. The degree of the rate increasing is function of the distance between the current *queue length* and the half of the buffer size. However, when the *queue length* is equal to the half of the buffer size and the *growth rate* is positive, the *flow rate* should be decreased moderately to save the *queue length* in this range.

This example shows again the flexibility provided by the fuzzy logic to control carefully the traffic adaptation. Such results cannot be obtained using a traditional weighted sum models.

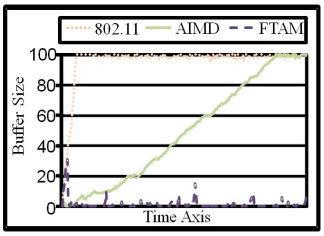
B. Simulation Results

In what follows, we present some preliminary simulation results of the proposed model using GlomoSim simulator. Throughout the simulation, each wireless node has a transmission range of 250 meters and shares an 11 Mbps radio channel with its neighboring nodes. The source and destination nodes associated with flows are distributed among the nodes in the wireless mesh network. The simulated environment has a square shape of 1000m x 1000m where all wireless nodes share a single radio channel of 11 Mbps. The performances of FTAM are compared with IEEE 802.11 and SWAN-AIMD [11]. During the simulation, real time voice and video flows are active and monitored. Voice and video traffic are modeled as 80 Kbps and 200 Kbps constant rate, respectively.

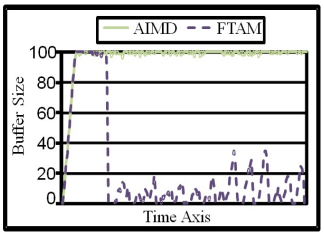
Figure 4a shows the buffer size variations at a source node at the beginning of the simulation. This figure shows clearly that the buffer size in FTAM is always limited. At the beginning of the simulation, the buffer size may go up to 30%, this is due essentially to the fact that there is no much traffic in intermediates nodes; hence it is possible to send additional packets over these nodes. Figure 4a illustrates also that buffers in AIMD remains almost full for a longer period of time waiting for successive messages losses. It is clear that FTAM improves exponentially the performance of 802.11 in terms of QoS. Therefore, it would be more significant to compare FTAM to an enhanced and well known approach as SWAN-AIMD.

Figure 4b is a capture of the buffer size variation at an intermediate node. We can observe clearly that FTAM attempts to reach the maximum tolerable throughput; which means that the buffers in intermediate nodes will be filled as rapidly as in SWAN-AIMD. Nevertheless, a best value of throughput is reached more quickly. Thus, the buffer size in intermediate nodes will take a small value which can help to reduce the congestion and to decrease the time delivery of traffic packets. Regarding the buffers in AIMD, they are filled gradually upon the detection of packets lost problem.

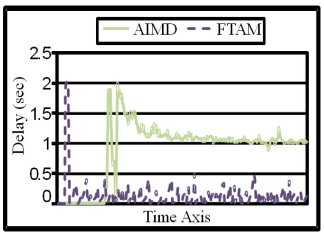
Figure 4c illustrates the impact of the buffers variations at intermediate nodes on the end-to-end delay. We observe that AIMD increments gradually the throughput upon the detection of congestion problem, and packets will wait longer in buffers in intermediate nodes. However, our model, FTAM, maintains the size of buffers as low as possible; thus the traffic packets do not lose much time, waiting in the intermediate nodes.



(a) Buffer size at a source node



(b) Buffer size at a intermediate node



(c) Delay

Fig. 4: Buffer size and delay variations

VI. Conclusion

This paper explored the usage of artificial intelligent (Fuzzy Logic) technique in order to control the rate adaptation of multimedia real time traffic in WMNs. The proposed model, FTAM, uses the queue length variation rate in addition to the current queue length in order to predict and control the

 ISBN: 978-1-61208-126-7

network congestion. One of the benefits of FTAM is that the regulation rate is predicted as soon as the congestion is expressed in the nodes. Moreover, FTAM ensures that best-effort traffic coexists well with real-time traffic in the multimedia applications.

Future works include extensive simulations using GlomoSim underboth the single-hop and multi-hop environment and compare the performances of FTAM with the standard IEEE 802.11. and SWAN-AIMD

REFERENCES

[1] M. Natkaniec, "Ad Hoc Mobile Wireless Networks: Principles, Protocols, and Applications," *IEEE Communications Magazine*, vol. 47, no. 5, pp. 12–14, May 2009.

[2] I. Akyildiz and X. Wang, "A Survey on Wireless Mesh Networks," *IEEE Communications Magazine*, vol. 43, no. 9, pp. S23–S30, 2005.

[3] X. Xiang, X. Wang, and Y. Yang, "Stateless Multicasting in Mobile Ad Hoc Networks," *IEEE Transactions on Computers*, vol. 59, no. 8, pp. 1076–1090, 2010.

[4] A. Kaufmann, *Introduction to Theory of Fuzzy Subsets*. New York: Academic, 1975.

[5] W. Pedrycz and F. Gomide, *An Introduction to Fuzzy Sets: Analysis and Design*. MIT Press, 1998.

[6] L. Zadeh, "Fuzzy Logic = Computing with Words," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 2, pp. 103–111, May 1996.

[7] L. Khoukhi and S. Cherkaoui, "Intelligent QoS Management for Multimedia Services Support in Wireless Mobile Ad Hoc Networks," *Computer Networks*, vol. 54, no. 10, pp. 1692–1706, 2010.

[8] Z.-X. Wang, H.-L. Xia, and W. Ding, "A Fuzzy Logic Cooperative MAC for MANET," *The Journal of China Universities of Posts and Telecommunications*, vol. 15, no. 1, pp. 55–60, May 2008.

[9] R. de Oliveira and T. Braun, "A Delay-based Approach using Fuzzy Logic to improve TCP Error Detection in Ad Hoc Networks," in *Proceedings of the IEEE WCNC Conference*, vol. 3, 2004, pp. 1666–1671.

[10] L. Reznik and V. Kreinovich, "Fuzzy and Probabilistic Models of Association Information in Sensor Networks," in *Proceedings of the IEEE Fuzzy Systems Conference*, vol. 1, 2004, pp. 185–189.

[11] G. Ahn, A. T. Campbell, A. Veres, and L. H. Sun, "SWAN: Service Differentiation in Stateless Wireless Ad Hoc Networks," in *Proceedings of the IEEE INFOCOM Conference*, vol. 2, 2002, pp. 457–466.

 ISBN: 978-1-61208-126-7

Measuring The Satisfaction Degree Of Quality Attributes Requirements For Services Orchestrations

Nabil Fakhfakh, Hervé Verjus, Frédéric Pourraz, and Patrice Moreaux
University of Savoie
LISTIC Laboratory
Annecy-Le-Vieux, France
Email: {nabil.fakhfakh, herve.verjus, frederic.pourraz , patrice.moreaux}@univ-savoie.fr

***Abstract*—Quality attributes are an important concern in critical business processes supported through services orchestration. Various works in the literature have addressed quality attributes values computation of services orchestration. This paper proposes a novel approach that combines workflow patterns aggregation rules and a multi-criteria decision making method named MACBETH. This approach allows us to measure the satisfaction degree of services orchestration to the quality attributes requirements as defined at design-time.**

Keywords-quality of service/quality attributes aggregation; SOA; satisfaction degree measurement.

I. INTRODUCTION

Service Oriented Architectures (SOAs) is an emerging paradigm for the development of business applications supported through services orchestrations. Services are the central concept of SOAs. Multiple services with same functionalities are spread over Internet. This makes clients faced to a large choice between functionally equivalent services as well as services providers faced to increased competition. Therefore, both clients and services providers need to agree on some guarantees: clients require guarantees that satisfy their expectations and services providers need guarantees to prevent from hazardous use of their services.

Today, quality attributes for services and services orchestrations become an important issue. At design-time, clients choose services satisfying services orchestration requirements in term of quality attributes (also referred as Quality of Service in the literature). However, at execution time, some services involved in the orchestration may unpredictably fail or change their quality attributes values positively or negatively. This will result in possible deviations of quality attributes values of the orchestration. Therefore, clients need to know how much the orchestration meets their requirements.

Various works dealing with computation of services orchestration quality attributes values exist in the literature [1] [2] [3]. Currently, none approach supports different kinds of quality attributes simultaneously in order to provide a high level information for decision making. The satisfaction degree of services orchestration is a such high level information. Measuring the satisfaction degree of services orchestrations can be seen as a Multi-Criteria Decision Making (MCDM) problem. In this paper, we present a novel approach giving a measurement of the satisfaction degree of services orchestrations. This approach has the advantage of taking into account clients preferences from quality attributes point of view. It is based on workflow patterns aggregation rules [1] [2] [3] and a MCDM method named the Measuring Attractiveness by a Categorical-Based Evaluation TecHnique (MACBETH) [4].

The remainder of this paper is organized as follows. Section II introduces related work on quality attributes values aggregation and MCDM methods. Section III details our approach for measuring services orchestration satisfaction degree, while Section IV concludes the paper.

II. RELATED WORK

Various approaches have been proposed to compute each quality attribute value independently for services orchestration (e.g., giving response time values for all services composing the orchestration, how to compute the overall response time of the orchestration?). These approaches can be classified in two categories: *probabilistic models-based approaches* [6][7][8][9] and *workflow patterns-based approaches* [1][2][3][5]. The first category of approaches that allow to compute quality attributes values of services orchestration is based on probabilistic models. It consists in transforming the services orchestration model into a probabilistic model (e.g., Continuous Time Markov Chain (CTMC) model [6][7] or Discrete Time Markov Chain (DTMC) [8] or Stochastic Petri Nets (SPN)[9]). Then, the probabilistic model is annotated with quality attributes values. Finally, these approaches use tools like PRISM [6] or SPNP [9] to compute each quality attribute value of services orchestration. The major drawback of these approaches is that they support only reliability and/or response time. The second category consists in defining aggregation rules of quality attributes values for each composition pattern. A composition pattern is a combination of pairs of workflow patterns [10]; it is composed from one split pattern

 ISBN: 978-1-61208-126-7

Table I: Workflow Pattern-Based Approaches and Related Works

Research work	composition patterns	quality attributes
Jaeger and al. [2]	Sequence, loop, AND-AND, XOR-XOR, AND-XOR, OR-OR, OR-XOR, AND-N/M, OR-N/M	throughput, response time, cost, availability, reputation, encryption grade
Rosenberg and al. [3]	Sequence, loop, AND-AND, XOR-XOR	throughput, response time, cost, availability, reputation, encryption grade, scalability, accuracy, robustness
Cardoso and al. [1]	sequence, loop, AND-AND, XOR-XOR, Fault-tolerant systems	response time, cost, reliability, fidelity
Coppolino and al. [5]	sequence, loop, AND-AND, XOR-XOR, AND-N/M	reliability

(e.g., AND-split) and one join (synchronisation) pattern (e.g., XOR-join) except the sequence and loop patterns, which are considered individually. Table I summarizes, for each research work, the composition patterns taken into account and the supported quality attributes. The advantage of workflow patterns-based approaches is that they support larger set of quality attributes. Moreover, they are extensible: *(i)* more composition patterns could be added and *(ii)* new quality attributes could be integrated. For that reason, in our proposal, we will exploit a workflow patterns-based approach.

However, when changes affect positively and/or negatively some quality attributes values, it becomes difficult to estimate how much the whole orchestration fits client' expectations and satisfaction. Thus, it would be useful to have a high-level information. This information represents the services orchestration satisfaction degree. This could be done by aggregating the quality attributes values to provide only one value that measures the satisfaction degree of the services orchestration. To this end, we propose to use a MCDM method.

The most common MCDM method used in industrial applications is Analytical Hierarchy Process (AHP) [11]. It is based on the Weighted Arithmetic Mean (WAM) operator. AHP allows the elementary performance expression and the WAM operator weights quantification using human expertise. However, this method suffers from a lack of consistency between the expression of elementary performances step and the determination of the weights step: the weights being expressed on a ratio scale are not consistent with the interval scale of the elementary performances.The MACBETH method [4], as the AHP method, defines quantitative performance expression and aggregation from qualitative pairwise comparisons of situations issued from the decision-maker. But unlike AHP, the MACBETH methodology satisfies the measurement theory requirements (i.e., with respect to the commensurability and significance [4]). That's why, we choose the MACBETH method to aggregate quality attributes values and to measure the satisfaction degree of the services orchestration.

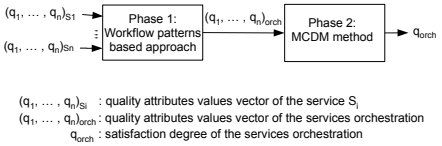


Figure 1: Principe of the Aggregation Approach

III. AGGREGATION PROPOSAL

In this section, we introduce our proposal for measuring the satisfaction degree of services orchestrations. We define the satisfaction degree as the percentage of respect to quality attributes requirements specified at design time (e.g., in SLAs). In this paper, we do not deal with requirements definition and we assume that these requirements are provided.

Each service S_i, involved in a given services orchestration, has a quality attributes values vector (see Fig. 1) denoted as: $(q_1, ..., q_n)_{Si}$, where $q_{1 \leq j \leq n}$ is the quality attribute value. Starting from these vectors, they are firstly aggregated in phase 1 using workflow patterns aggregation rules. This results in one vector of quality attributes values of the orchestration (see Fig. 1). Next, the quality attributes values of this vector will be aggregated in phase 2 using a MCDM method. That provides us only one value, which is the satisfaction degree of services orchestration. Before detailing these two phases, we present a simplified manufacturing process supported through a services orchestration described in Fig. 2. We will use this services orchestration model to illustrate our approach hereafter.

The process begins with launching a production request (service operation *LaunchProductionRequest*). Then, when the production terminates, the production planning is updated (service operation *UpdateProductionPlanning*); in parallel, a quality control is executed on the product (service operation *LaunchQualityControl*). If the control is positive, the product is moved to stock (service operation *MoveProductToStock*), otherwise it will be reported as wasted product (service operation *Report WasteProduct*).

Hereafter, we assume that quality attributes requirements

 ISBN: 978-1-61208-126-7

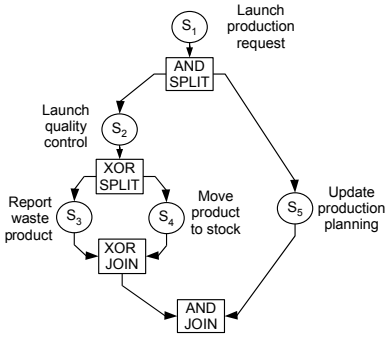


Figure 2: Example of Manufacturing Process

and measurements are given respectively by the client and a monitoring system (quality attributes measurement is out of the scope of this paper).

A. Phase 1: Aggregation Based on Workflow Patterns Rules

In the first phase (see Fig. 1), we use aggregation rules based on workflow patterns (see Section II) to compute each quality attribute value of the services orchestration. This consists in applying step-by-step rules in order to aggregate quality attributes values. The applied rules are those corresponding to the composition patterns used in the services orchestration model. Beginning from the most nested composition pattern, the orchestration model is parsed and aggregation rules for each value in the quality attributes values vector are progressively applied. This is done until reducing the whole orchestration model into a sole node (Fig. 3). The resulted quality attributes values of the end node form the quality attributes values vector of the services orchestration. This approach is relevant for each quality attribute value that has aggregation rules for the composition patterns. We will detail this phase through the services orchestration described in Fig. 2.

For simplification purpose, we will consider a set of three quality attributes values: *response time* (q_{rt}), *reliability* (q_{rel}) and *availability* (q_{av}). The aggregation rules for each of the composition patterns and for each of the quality attributes values are summarized in Table II. The first step consists in checking the most nested composition pattern, which is the XOR-XOR pattern between S_3 and S_4 in Fig. 3a. Then, we apply the respective aggregation rule from Table II. The quality attributes values computation of this composition pattern gives :

$$q_{rt}(S_{3,4}) = p_1.q_{rt}(S_3) + p_2.q_{rt}(S_4)$$
$$q_{rel}(S_{3,4}) = p_1.q_{rel}(S_3) + p_2.q_{rel}(S_4)$$
$$q_{av}(S_{3,4}) = p_1.q_{av}(S_3) + p_2.q_{av}(S_4)$$

Thus, the orchestration model is reduced to that given in Fig. 3b. Then, taking into account the reduced orchestration model, the next composition pattern to be considered is the sequence pattern of S_2 and $S_{3,4}$. The quality attributes values computation of this composition pattern gives :

$$q_{rt}(S_{2,3,4}) = q_{rt}(S_2) + q_{rt}(S_{3,4})$$
$$q_{rel}(S_{2,3,4}) = q_{rel}(S_2).q_{rel}(S_{3,4})$$
$$q_{av}(S_{2,3,4}) = q_{av}(S_2).q_{av}(S_{3,4})$$

The following composition pattern identified in step 3 (Fig. 3c) is the parallel pattern associated to the synchronization pattern (AND-AND). The resulting quality attributes values after the reduction of the orchestration model are:

$$q_{rt}(S_{2,3,4,5}) = \max(q_{rt}(S_{2,3,4}), q_{rt}(S_5))$$
$$q_{rel}(S_{2,3,4,5}) = q_{rel}(S_{2,3,4}).q_{rel}(S_5)$$
$$q_{av}(S_{2,3,4,5}) = q_{av}(S_{2,3,4}).q_{av}(S_5)$$

The obtained orchestration model from this step is composed of two nodes structured in sequence (Fig. 3d). By aggregating quality attributes values of these two nodes in sequence, the services orchestration is reduced to a sole node. The quality attributes values of this node form the quality attributes values vector of the whole services orchestration model. This vector is denoted as $(q_{rt}(orch), q_{rel}(orch), q_{av}(orch))$ where :

$$q_{rt}(orch) = q_{rt}(S_{1,2,3,4,5}) = q_{rt}(S_1) + q_{rt}(S_{2,3,4,5})$$
$$q_{rel}(orch) = q_{rel}(S_{1,2,3,4,5}) = q_{rel}(S_1).q_{rel}(S_{2,3,4,5})$$
$$q_{rel}(orch) = q_{av}(S_{1,2,3,4,5}) = q_{av}(S_1).q_{av}(S_{2,3,4,5})$$

This resulting vector $(q_{rt}(orch), q_{rel}(orch), q_{av}(orch))$ or more simply $(q_{rt}, q_{rel}, q_{av})_{orch}$ will be the input of the phase 2 (Fig. 1).

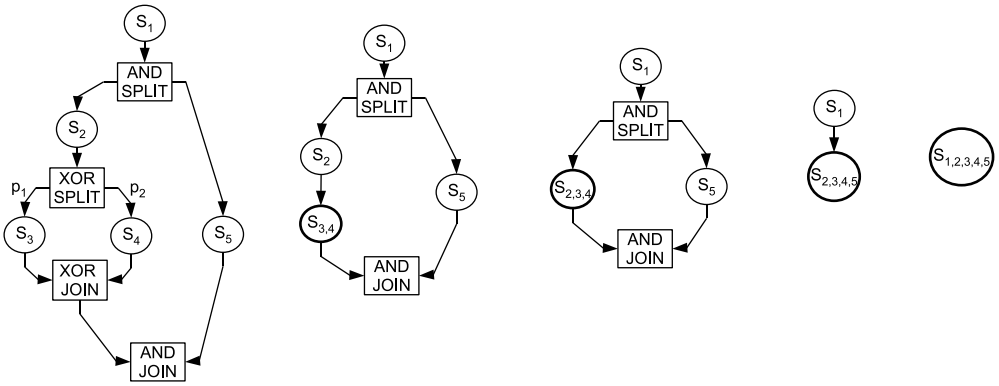
B. Phase 2: Aggregation Based on Weighted Mean Method (MACBETH)

The goal of this phase is to aggregate different values in the quality attributes values vector of the services orchestration (e.g., $(q_{rt}, q_{rel}, q_{av})_{orch}$) in order to obtain a measure of the satisfaction degree of the services orchestration (q_{orch}) (see Fig. 1). This measure allows us to interpret the positive or negative changes that affect the quality attributes values vector of the orchestration. For example, how to compare the vectors $(10, 0.95, 0.99)_{orch}$ and $(7, 0.90, 0.99)_{orch}$. A client who has strong time constraint may consider variation of the quality attributes values as an improvement of quality orchestration. An other one who is interested to reliability can report this variation as a degradation. For that purpose and in order to discriminate these alternatives, we use the MACBETH method [4]. MACBETH is a weighted mean

Table II: Aggregation Rules [1][3]

	Response Time	Reliability	Availability
Sequence	$\sum_{i=1}^{n} q_{rt}(s_i)$	$\prod_{i=1}^{n} q_{rel}(s_i)$	$\prod_{i=1}^{n} q_{av}(s_i)$
Loop	$q_{rt}(s_i) * c$	$q_{rel}(s_i)^c$	$q_{av}(s_i)^c$
AND-AND	$\max(q_{rt}(s_1), ..., q_{rt}(s_n))$	$\prod_{i=1}^{n} q_{rel}(s_i)$	$\prod_{i=1}^{n} q_{av}(s_i)$
XOR-XOR	$\sum_{i=1}^{n} p_i.q_{rt}(s_i)$	$\sum_{i=1}^{n} p_i.q_{rel}(s_i)$	$\sum_{i=1}^{n} p_i.q_{av}(s_i)$

[a] c denote the number of occurring loops
[b] p_i the probabilities of the outgoing branches for XOR-XOR



(a) Step 1 (b) Step 2 (c) Step 3 (d) Step 4 (e) Step 5

Figure 3: Workflow Pattern-Based Aggregation Steps

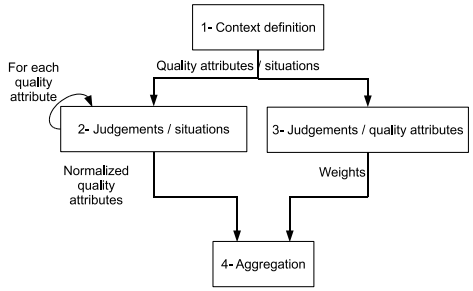


Figure 4: The Main Steps of MACBETH Method

method that allows us to translate quality attributes values to a satisfaction degree and thus thanks to :

- normalization of quality attributes values according to measurement theory (i.e., with respect to the commensurability and significance requirement [4]),
- determination of the weights of the weighted mean operator,
- aggregation of normalized quality attributes values.

It is based on pairwise comparisons of situations made by the client (the expert). The MACBETH method comprises four main steps (Fig. 4):

1) *Context definition step:* The first step consists in identifying the criteria, which are quality attributes in our case. Secondly, situations that will be compared are defined. In our context, situations are represented by vectors of quality attributes values that result from several instantiations of the services orchestration model. For example, we consider three instances of services orchestration represented by measured quality attributes values as follows:

$$S^1 = (q^1_{rt}, q^1_{rel}, q^1_{av}) = (10, 0.9, 1)$$
$$S^2 = (q^2_{rt}, q^2_{rel}, q^2_{av}) = (15, 0.6, 0.8)$$
$$S^3 = (q^3_{rt}, q^3_{rel}, q^3_{av}) = (5, 1, 0.6)$$

Two more reference situations are introduced denoted situation *good* and situation *neutral*. The *good* situation represents the total satisfaction of the requirements associated to the criteria (i.e., the best quality attributes values vector) while the *neutral* situation represents the minimum satisfaction of the requirements (i.e., the worst quality attributes values vector). These two reference situations could be retrieved for example from Service Level Agreement (SLA) contracts. For

 ISBN: 978-1-61208-126-7

example :

$$S^{good} = (q_{rt}^{good}, q_{rel}^{good}, q_{av}^{good}) = (3, 1, 1)$$
$$S^{neutral} = (q_{rt}^{neutral}, q_{rel}^{neutral}, q_{av}^{neutral}) = (20, 0.5, 0.5)$$

Note that the *good* situation and the *neutral* situation are associated respectively to the vectors of elementary performance expression $(1, 1, 1)$ and $(0, 0, 0)$ (i.e., after normalization). Therefore, if we consider that these two situations are formed from upper and lower bounds of quality attributes values, all other situations will be classified between them.

2) *The elementary performance expression step:* In this step, the goal is to normalize quality attributes values. To this end, the client (the expert) uses his expertise to judge given situations and fulfill the matrix of judgements like the one given in Table III. Firstly, he is asked for each criterion (i.e., quality attribute) about his preferences between pairs of situations (including the two reference situations). If the client prefers situation S^i to S^j for a criterion k, this is noted as follows:

$$S^i \succ S^j$$

and means that for the normalized quality attributes values $q_k^i > q_k^j$. This is mapped in Table III into the classification of the situations by their order of preference.
Secondly, the client expresses his strengths of preference about the same situations. The strengths of preference are characterized with seven levels: *0=null, 1=very weak, 2=weak, 3=moderate, 4=strong, 5=very strong, 6=extreme* (see Table III). If the client cannot give his strengths of preference but only his preferences, this is noted by *positive* or more shortly *P*. The client prefers the situation S^i to S^j with a strength $h \in \{0, ..., 6\}$ for a criterion k i.e.,

$$S^i \succ^{\mathbf{h}} S^j$$

This is equivalent to :

$$q_k^i - q_k^j = \mathbf{h}\alpha$$

where α is a coefficient necessary to meet the condition q_k^i and $q_k^j \in [0, 1]$.
Example : We ask the client to compare the response time of the situations S^1 and S^2. He says that he prefers S^1 to S^2 i.e.,

$$S^1 \succ S^2 \Leftrightarrow q_{rt}^1 > q_{rt}^2$$

and that he prefers extremely $(h = 6)$ S^1 to S^2 i.e.,

$$S^1 \succ^6 S^2 \Leftrightarrow q_{rt}^1 - q_{rt}^2 = 6\alpha$$

Next, once all the strengths of preference between situations are provided. The matrix of judgements is fulfilled (e.g., see Table III) and a system of equations can be extracted. It takes the form:

$$q_k^i - q_k^j = h\alpha$$

By solving this system of equations, elementary performance expressions (normalized quality attributes values) are quantified in the interval [0,1].
Example: for the strengths of preference expressed in Table III, the system of equations is the following :

$$\begin{aligned} (q_{rt}^{good} = 1)\ q_{rt}^{good} - q_{rt}^3 = 1 - q_{rt}^3 &= 4\alpha \\ q_{rt}^3 - q_{rt}^1 &= 5\alpha \\ q_{rt}^1 - q_{rt}^2 &= 6\alpha \\ (q_{rt}^{neutral} = 0)\ q_{rt}^2 - q_{rt}^{neutral} = q_{rt}^2 - 0 &= 6\alpha \end{aligned}$$

The above system is solvable and the solution is:

$$q_{rt}^1 = 0.5714,\ q_{rt}^2 = 0.2857,\ q_{rt}^3 = 0.8095$$

Note that the same procedure is established for each quality attribute (i.e., for reliability and availability too). In the same way, we get for reliability and availability:

$$q_{rel}^1 = 0.6364,\ q_{rel}^2 = 0.0909,\ q_{rel}^3 = 1$$
$$q_{av}^1 = 1,\ q_{av}^2 = 0.7333,\ q_{av}^3 = 0.4$$

3) *Weights determination step:* MACBETH is based on the weighted mean operator for the aggregation given as follows:

$$q_{ag} = \sum_{i=1}^{n} w_i . q_i \qquad (1)$$

where w_i represents the relative importance of quality attribute q_i. Thus, to determine the weights w_i, the client expresses his judgements about the relative importance of quality attributes. To this end, he has to compare particular (possibly fictive) situations where only one normalized quality attribute value is set to 1 e.g., (0,...,0,1,0,...,0). So, the aggregated quality attribute value is reduced to $q_{ag} = w_i$, where q_{ag} is the aggregated value of the quality attributes values vector

Table III: Preferences and Preferences Strengths for Response Time

Response Time	Good	S^3	S^1	S^2	Neutral
Good	No	strong [a]	P	P	P
S^3		No	very strong	P	P
S^1			No	extreme	P
S^2				No	extreme
Neutral					No

[a] 0=null, 1=very weak, 2=weak, 3=moderate, 4=strong, 5=very strong, 6=extreme

Table IV: Expert Judgements for Weights Determination

	$(0,1,0)$	$(0,0,1)$	$(1,0,0)$	$(0,0,0)$
$(0,1,0)$	No	Strong	Positive	Positive
$(0,0,1)$		No	Very weak	Positive
$(1,0,0)$			No	Very weak
$(0,0,0)$				No

with $q_i = 1$ and $q_j = 0$ for $i \neq j$. To determine the *n* weights of the quality attributes, the client has to give at least *n* strengths of preference between considered situations (e.g., Table IV). This results in *n* equations:

$$q_{ag}^{i} - q_{ag}^{j} = h\alpha = w_i - w_j$$

with the condition $\sum_{i=1}^{n} w_i = 1$. Note that $h \in \{0,...6\}$ represents the strengths of preference.
Example: let us consider a quality attributes values vector with three quality attributes values $(q_{rt}, q_{rel}, q_{av})$. So we have to compute three weights w_{rt}, w_{rel} and w_{av}. Taking the matrix of judgements fulfilled by the client (Table IV), we can write the following equations :

$$\begin{aligned} q_{ag}^{(0,1,0)} - q_{ag}^{(0,0,1)} &= 4\alpha = w_{rel} - w_{av} \\ q_{ag}^{(0,0,1)} - q_{ag}^{(1,0,0)} &= \alpha = w_{av} - w_{rt} \\ q_{ag}^{(1,0,0)} - q_{ag}^{(0,0,0)} &= \alpha = w_{rt} - 0 \\ w_{rt} + w_{rel} + w_{av} &= 1 \end{aligned}$$

The solution of the system of equations is:

$$w_{rt} = 0.1111,\ w_{rel} = 0.6666,\ w_{av} = 0.2223$$

4) *Aggregation step:* In this step, as quality attributes values are normalized (step 2) and weights are computed (step 3), we have just to apply the formula (1) for every situation characterized by a quality attributes values vector $(q_1, ..., q_n)$.
Example: The aggregated values of situations S^1, S^2 and S^3 are :

$$q_{ag}^{S^1} = 0.71\ , q_{ag}^{S^2} = 0.2553\ , q_{ag}^{S^3} = 0.8455$$

The aggregation results show that the situation S_3 is better than S_1 and that S_1 is better than S_2.

IV. Conclusion and Future Work

In this paper, we have proposed a novel approach for measuring the satisfaction degree of services orchestrations. Our work aims to preserve services consumer satisfaction and to do the necessary adaptations whenever any violation of the services consumer requirements occurs. However, as seen in Section III-B, the reference situations (situations *Good* and *Neutral*) are defined respectively by the best quality attributes values and by the minimum ones agreed by the client. Therefore, the minimum satisfaction degree that can be measured (i.e., equal to "0") correspond to the minimum quality attributes values agreed by the client. So, the approach does not allow us to predict possible violation of client requirements. This will be the subject of our future work.

Acknowledgements

This work is partially funded by the FEDER MES project granted by the French Rhône-Alpes Region.

References

[1] J. Cardoso, J. Miller, A. Sheth, and J. Arnold, "Modeling quality of service for workflows and web service processes," *Journal of Web Semantics*, vol. 1, pp. 281–308, 2002.

[2] M. C. Jaeger, "Optimising quality-of-service for the composition of electronic services," Ph.D. dissertation, Berlin University, Germany, 2007.

[3] F. Rosenberg, "Qos-aware composition of adaptive service-oriented systems," Ph.D. dissertation, Technical University Vienna, Austria, June 2009.

[4] V. Cliville, L. Berrah, and G. Mauris, "Quantitative expression and aggregation of performance measurements based on the macbeth multi-criteria method," *International Journal of Production Economics*, vol. 105, pp. 171–189, 2007.

[5] L. Coppolino, L. Romano, N. Mazzocca, and S. Salvi, "Web services workflow reliability estimation through reliability patterns," *Security and Privacy in Communications Networks and the Workshops,*, pp. 107–115, 2007.

[6] S. Gallotti, C. Ghezzi, R. Mirandola, and G. Tamburrelli, "Quality prediction of service compositions through probabilistic model checking," in *QoSA '08: Proceedings of the 4th International Conference on Quality of Software-Architectures*, Berlin, Heidelberg, 2008, pp. 119–134.

[7] N. Sato and K. S. Trivedi, "Stochastic modeling of composite web services for closed-form analysis of their performance and reliability bottlenecks," in *ICSOC '07: Proceedings of the 5th international conference on Service-Oriented Computing*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 107–118.

[8] V. Cortellessa and V. Grassi, "Reliability modeling and analysis of service-oriented architectures," in *Test and Analysis of Web Services*, 2007, pp. 339–362.

[9] D. Zhong and Z. Qi, "A Petri net based approach for reliability prediction of web services," in *OTM Workshops (1)*, 2006, pp. 116–125.

[10] W. van der Aalst, A. ter Hofstede, B. Kiepuszewski, and A. Barros, "Workflow patterns," in *Distributed and Parallel Databases*, vol. 14 (1), 2003, pp. 5–51.

[11] E. H. Forman and M. A. Selly, *Decision by objectives: how to convince others that you are right*. World Scientific Pub Co Inc, 2001.

 ISBN: 978-1-61208-126-7

A TDMA-based MAC Protocol for Wireless Mesh Networks using Directional Antennas

Ali El Masri, Lyes Khoukhi, and Dominique Gaiti
ICD/ERA, Troyes University of Technology
12 rue Marie Curie, 10010 Troyes Cedex , France
Email: {ali.el_masri, lyes.khoukhi, dominique.gaiti}@utt.fr

***Abstract*—Wireless Mesh Networks (WMNs) have emerged as a key technology for next generation wireless networks. It is expected that WMNs will support diverse kind of multimedia services. Hence, there is a vital need to conceive efficient and reliable MAC protocols that allows alleviating wireless MAC problems. In this paper, we develop a novel MAC scheme for Wireless Mesh Networks based on a TDMA approach and using directional antennas. Our main objective behind this proposal is to respond to the main directional MAC problems such as deafness and hidden terminal problems caused by the lack in knowledge at each wireless node about its neighboring transmissions. The contribution of our MAC scheme is twofold. First, in enabling parallel transmissions and second in maintaining a full knowledge at each wireless node about neighboring transmissions. Using parallel transmissions, a busy node will not miss the start of a neighboring transmission. A robust algorithm of mini-slot assignment is proposed to provide full-knowledge, collision-free, and a fair channel access between all the nodes in the network.**

***Index Terms*—Wireless Mesh Networks; Directional Antennas; Medium Access Control (MAC);**

I. INTRODUCTION

Wireless mesh networks (WMNs) have emerged as a key technology for next-generation wireless networks. A WMN is a set of stationary wireless routers serving as access points for wireless clients and forming together a wireless mesh backbone (creating in effect, an ad hoc network).

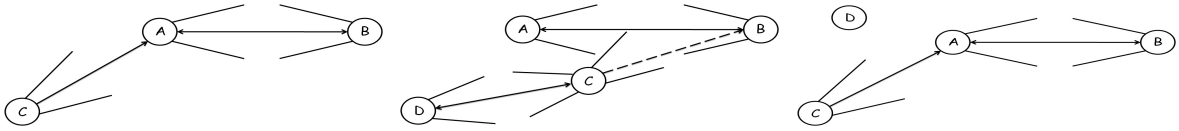
One of the main challenges that are facing the deployment of WMN is how to ensure Quality of Service (QoS) to the real-time traffic. In fact, a properly designed MAC protocol is a key factor to maintain the reliability of transmissions, to satisfy the QoS requirements and to efficiently allocate the radio resources. Two main aspects distinguish WMNs from traditional multi-hop wireless networks such as mobile ad hoc and wireless sensor networks [1]. First, the routers are placed in fixed location with continuous power supply, and then there is no mobility or energy constraints. Second, the traffic aggregated at each local router is heavy due to the forwarded traffic from neighboring routers and the uploaded traffic from local wireless clients, and then contention-based MAC protocols suffer from serious collisions and great violations of QoS requirements. Hence, designing a properly MAC protocols continues to be critical problem in WMNs.

In the other hand, directional antennas have been successfully applied to increase spatial reuse, reduce interference, extend transmission range and optimize power consumption. Directional antennas seem then to be an appropriate technique to improve the performance of WMNs characterized by their heavy load traffic and their fixed wireless routers. The past few years have witnessed the special attention given by the wireless community to the directional antennas concept. A number of proposed MAC protocols based on directional antennas have been proposed, most of them are CSMA/CA-based protocols. However, the use of directional antennas introduces new MAC problems such as deafness problem, new kinds of hidden terminal problems and head-of-line blocking problem.

The deafness problem (Figure 1a) occurs when a node C tries to communicate with a node A while node A is beamformed to another direction than the direction of node C. In such scenario, node C loses many RTS (Ready-To-Send) packets while trying to establish the connections with A. It leads to the following: Increase the contention window of node C, lose the fairness between nodes, waste the bandwidth and cancel the spatial reuse, which it is the main benefit of directional antennas. The hidden terminal problem (Figure 1b) occurs when two nodes A (source) and B (destination) begin a communication while a neighboring node C is already communicating with other nodes. When node C becomes idle, it does not sense the communication between A and B due to the use of directional antennas. If node C has in the head of its queue a packet intended to node B, it starts the transmission of a series of unsuccessful RTS packets. Node C suffers then from the inconveniences of deafness problem explained above, and node B suffers from collisions between data packets received from A and RTS packets received from C. The head-of-line problem (Figure 1c) occurs when a node C has in the head of its queue a packet intended to a busy node A, while the next packet in the queue is intended to an idle node D. Hence, the packet intended to D suffers from unnecessary delay.

The main reason of these problems is the lack in awareness at a wireless node about the concurrent neighboring transmissions. This lack is due to one of the following two scenarios. First, when two nodes begin a communication, they do not inform all the nodes in their vicinity about the ongoing communication due to the use of directional RTS/CTS (Clear-To-Send) transmissions. Second, even if nodes in the vicinity of the communication are well informed, the RTS/CTS handshake cannot be heard by a busy node. In our opinion, parallel transmissions seem to be an optimal solution of these problems. We mean by parallel transmissions that the different concurrent transmissions have the same start and end times.

 ISBN: 978-1-61208-126-7



(a) Deafness problem (b) Hidden terminal problem (c) Head-of-line blocking problem

Fig. 1: Directional MAC problems

In this paper, we propose a new TDMA-based (Time Division Multiple Access) MAC protocol for wireless mesh backbone using directional antennas. Our proposal aims to handle directional MAC problems by enabling parallel transmissions. The proposed protocol provides a full knowledge about the concurrent transmission in the surrounding of each wireless router. Thus, the proposed protocol is designed to be collision-free, deafness-free and blocking-free. In addition, the proposed MAC protocol includes a new algorithm that offers a flexible and scalable way for assigning time-slots to wireless routers. The proposed time-slot assignment algorithm is applied at each node/neighborhood, to completely avoid the transmission of unsuccessful control messages, and hence to widely reduce the control overhead in WMNs. Moreover, the time-slot assignment changes dynamically to ensure a fair channel access. Benefiting from the transmission characteristics of directional antennas, the spatial reuse is greatly increased by our protocol.

The rest of the paper is organized as follows. The related work is reviewed in Section II. The antennas model is presented in Section III. In Section IV, we detail our proposed protocol. Finally, we draw our conclusion in Section V.

II. Related Work

Several proposed MAC protocols for ad hoc and mesh networks using directional antennas exist in the literature. Most of them are based on the IEEE 802.11 DCF MAC protocol. In this section, we give our proper classification of MAC protocols using directional antennas. In our classification, protocols are divided into two groups: protocol with parallel transmissions and protocol with non-parallel transmission, each group is divided into several categories.

A. Non-Parallel Transmission

This group includes three categories of protocols, pure directional RTS/CTS, circular directional RTS/CTS, and explicit control messages category.

In Pure Directional Category [2][3], all RTS/CTS/Data/Ack packets are transmitted in the directional mode. Since no additional control packets exists, the spatial reuse is improved. However any proposed mechanism doesn't handle the deafness or the hidden terminal problem.

In the circular directional category [4][5], RTS and CTS are circularly transmitted in the directional mode through all or part of the unblocked beams of the transmitter and the receiver. The goal is to alleviate the deafness problem and to inform all the directional-omni neighbors about the ongoing transmission. Although the deafness problem is largely alleviated with this category of protocols, the deaf nodes are not completely avoided. In addition, the control overhead is largely increased due to additional control packets (circular RTS/CTS), and further fields inserted into RTS/CTS packets.

In the category of explicit control messages, additional control messages are used, e.g., busy-tone signals [6] and receiver initiated messages [7]. The busy-tone signals [6] are sent after the transmission to notify nodes in deafness state to reset their contention window. RTR (Ready To Receive) packet is sent by a receiver to invite potential sender to begin the transmission. The mean weakness of these approaches is that they try to solve directional MAC problems at the end of transmissions. The fairness is improved, however the unsuccessful RTS packets and bandwidth wastage still exist in this kind of protocols.

B. Parallel Transmission

Few works are proposed in the group of parallel transmission. These protocols are divided into two categories: CSMA-based [8], and TDMA-based [9]. In former case, the source-destination couple of the first successful exchange of RTS/CTS packets are considered as master nodes. Nodes in their vicinity are considered as slave nodes. After the successful exchange of control packets, master nodes wait for a duration before the beginning of transmission. Only the slave nodes, that win to exchange successful RTS/CTS packets during the waiting time, are allowed to transmit in parallel with the master nodes. Several limitations are observed: First, RTS/CTS packets are sent omnidirectionaly, hence the network connectivity is limited to OO-neighbors. Second, serious collisions may occur during the waiting time after the RTS/CTS exchange between the master nodes. Third, it is not clear how it will be the behavior of slave nodes if they are common neighbors of two couples of master nodes. To cope these issues, a directional TDMA-based MAC (RT-DMAC) protocol is proposed in [9]. However, RT-DMAC is designed only for uniform grid topologies where each node has four neighbors and is equipped with four beam antennas. The connectivity of RT-DMAC is also limited to OO-neighbors. This approach in not realistic in real WMNs and then the proposed protocol is not applicable.

III. Antennas Set-up

Each wireless router is equipped with M non-overlapped directional antennas. Routers can transmit or receive using directional or omnidirectional mode. When using omnidirectional mode, a router can either transmit or receive through all directions at the same time with a gain of G_o. When

 ISBN: 978-1-61208-126-7

using directional mode, a router can either transmit or receive through only one direction with a gain of $G_d \succ G_o$. Three sets of neighbors are used in our model: *OO-neighbors/DD-neighbors* when both the transmitter and the receiver use omnidirectional/directional mode respectively, and DO-neighbors when either the transmitter or the receiver uses the directional mode.

IV. The proposed MAC Protocol

A. Time Division

The time sequence in nodes is divided into slots. Each slot is divided into a control part and a transmission part. The goal of the control part is to decide which nodes will transmit in the following transmission part and which nodes will receive. In order to minimize the control overhead, the transmission part is much greater than the control part. The control part is divided into N mini-slot when N represents the number of nodes in the most load two-hop neighborhood. Each mini-slot is divided into three mini-parts. The two-hop neighborhoods of a given node consist of all nodes containing in the one-hop or the two-hop neighbors of this node. The time division is shown in Figure 2.

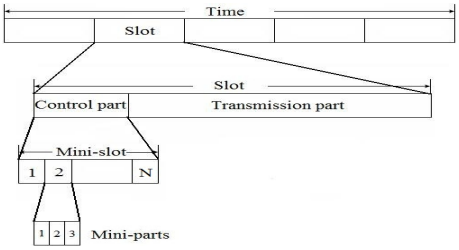


Fig. 2: Mini-slot Assignment

B. Mini-Slot Assignment

The goal of the mini-slot assignment is to assign a number of nodes to each mini-slot. If the number of mini-slots is small, then the control overhead is small. Hence, the number of nodes into each mini-slot should be the largest as possible. The mini-slot assignment should take into consideration that two nodes in the same two-hop neighborhood cannot be assigned to the same mini-slot. As explained in Section III-C, a node wishing to transmit during the transmission part should send a jam signal during it corresponding mini-slot. Hence, in order to maintain a collision-free MAC protocol, two nodes in the same two-hop neighborhood cannot be assigned to the same mini-slot. In fact, if two nodes in the same two-hop neighborhoods begin a transmission at the same time, resulting in a situation where the receiver of one of them is a one-hop neighbor of the other, and then a collision will occur at this receiver. The algorithm of mini-slot assignment is shown in Algorithm 1.

Algorithm 1 Mini-slot assignment

```
N ← 1; Number of mini-slots
L_i = The set of 2-hop neighbors of node i
S_j = The set of nodes in the mini-slot j
for i = 0 to Numberofnodes do
    flag ← false;
    for j = 0 to N do
        if L_i ∩ S_j == ø then
            add i to S_j;
            flag ← true;
            break;
        end if
    end for
    if flag == false then
        N ← N + 1;
        add i to S_N;
    end if
end for
```

C. Framework

At the beginning of each mini-slot, if a node S has a packet to transmit to node R and if S is assigned to this mini-slot, S will transmit a jam signal omnidirectionally during the first mini-part of the mini-slot. A jam signal is a busy-tone signal and it does not contain any information. The jam signal should be only sensed at the receiver but not decoded. When a receiver node R receives a jam signal during the first mini-part, it detects the direction of the received signal and it knows that the corresponding neighbor will be busy during the following transmission part. Note that since a jam signal only needs to be detected and not decoded, it will be sensed by all the nodes in the set of DO-neighbors of the sender S even if it is transmitted using the omnidirectional mode [10]. In the second mini-part, the node S transmits a jam signal toward the direction of R using the directional mode. Hence, when a neighbor node of S senses only the first jam signal, it knows that S will be busy during the following transmission part. In the other case, when a node receives a jam signal during the first and the second mini-slots, it knows that it will be the receiver node R during the following transmission part. In the result, all the one-hop neighbors of S are notified that S will be busy in the following transmission part. If one of the neighbors of S has in the head of its queue a packet intended to S, the transmission of this packet is deferred to the next slot. In the third mini-slot, the receiver R transmits a jam signal using the omnidirectional mode, and then the one-hop neighbors of R know that R will be busy during the following transmission part. In the following mini-slots, the assigned nodes can follow the same procedure if their corresponding receivers are not busy. In other words, the assigned nodes can follow this procedure if they have not received a jam signal from their corresponding receivers during the precedent mini-slots of the same slot. It is clear that using our protocol, all the nodes will have a complete knowledge about the concurrent transmission in their vicinity, and then no deafness, hidden, or blocking problem will take place since transmissions occur in parallel.

In the aim of better understating our proposed work, let's

 ISBN: 978-1-61208-126-7

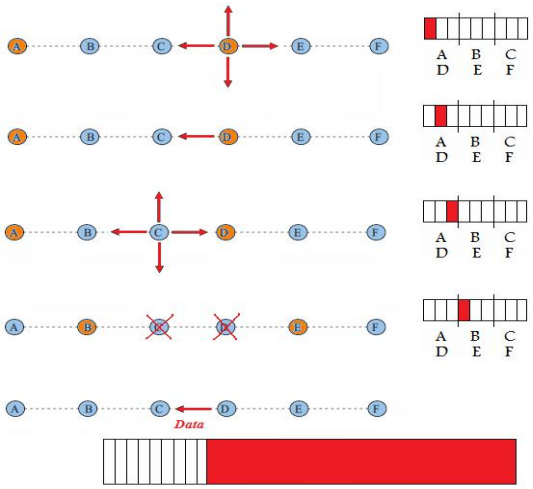


Fig. 3: The framework of our protocol

consider the example shown in Figure 3. It presents a chain topology of 6 nodes. In the first step, the model has to assign nodes to mini-slots. Let's consider the set of nodes A,B,C, since each node of this set is within the two-hop neighborhoods of the two other nodes, each one of them should be assigned to a separate mini-slot. Node D is not within the two-hop neighborhood of node A, and then node D is assigned to the same mini-slot of node A. For the same reasons, node E is assigned to the mini-slot of node B and node F is assigned to the mini-slot of node C. In the second step, we explain the operations realized by our protocol. Let's consider that at the beginning of the slot, the node D has in the head of its queue a packet intended to node C. During the first mini-part of the first mini-slot, node D transmits a jam signal using the omnidirectional mode. By sensing the direction of the receiving jam signal, node C and node E know that node D will be busy during the transmission part. During the second mini-part of the first mini-slot, node D transmits a jam signal toward node C using the directional mode. Hence, node C knows that it is the corresponding receiver of node D, and node E knows that it is not the corresponding receiver of node D. In the third mini-part, node C transmits a jam signal using the omnidirectional mode. By sensing the direction of the receiving jam signal, node B knows that node C will be busy during the transmission part. Nodes B and E are assigned to the second mini-slot. During the first mini-slot if node B has in the head of its queue a packet intended to node A, it follows the same procedure of node D. In the other case, if node B has in the head of its queue a packet intended to node C, it defers the transmission of this packet to the next mini-slot and it checks if it has another packet in the queue indented to node A. If such packet exists, node B follows the procedure of node D; otherwise B remains idle during its mini-slot. Node E follows the procedure of node B by preventing the transmission toward node D. The same procedure is repeated in the third mini-slot. The transmission will take place just after the end of the control part. The precedent mechanism is repeated at the beginning of each slot.

D. Fairness

It is clear from the previous sections that the nodes assigned to first mini-slots have the higher priority to access the channel. Hence, in order to maintain the fairness in channel access, the mini-slot assignment is shifted at the beginning of each slot. In other words, if a node is assigned to the i^{st} mini-slot in the current slot, it will be assigned to the $(i-1)^{st}$ mini-slot during the next slot. For instance, the nodes assigned to the second mini-slot in the current slot, will be assigned to the first mini-slot in the next slot. Hence, the priority of these nodes is incremented by one index. Also, the nodes assigned to the first mini-slot in the current slot will be assigned to the latest mini-slot in the next slot. Hence, the priority of these nodes will be changed from higher priority to lower priority.

V. Conclusion and Future Work

In this paper we have proposed a new TDMA-based MAC protocol for WMNs using directional antennas. By providing a full knowledge for each node about its neighboring transmissions and by maintaining parallel communications, our protocol is able to alleviate the control access challenges encountering in many MAC protocols. Our protocol is collision-free, deafness-free, hidden-free and blocking-free. Our proposol also provides a full faireness between all the nodes in the network.

Our future works will be focused on implementing and deploying our model, and adding more functionalities such as congestion control and service differentiation. In addition, the performances evaluations of our work will be compared with some existing directional MAC protocols.

References

[1] P. Wang and W. Zhuang, "A collision-free mac scheme for multimedia wireless mesh backbone," *IEEE Transactions on Wireless Communications*, vol. 8, no. 7, pp. 3577 –3589, 2009.

[2] M. Takai, J. Martin, R. Bagrodia, and A. Ren, "Directional virtual carrier sensing for directional antennas in mobile ad hoc networks," in *Proceedings of the ACM MOBIHOC Conference*, 2002.

[3] R. R. Choudhury, X. Yang, R. Ramanathan, and N. H. Vaidya, "Using directional antennas for medium access control in ad hoc networks," in *Proceedings of the ACM MOBICOM Conference*, 2002, pp. 59–70.

[4] T. Korakis, G. Jakllari, and L. Tassiulas, "Cdr-mac: A protocol for full exploitation of directional antennas in ad hoc wireless networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 2, pp. 145 –155, 2008.

[5] P. Li, H. Zhai, and Y. Fang, "Sdmac: Selectively directional mac protocol for wireless mobile ad hoc networks," *Wireless Networks*, vol. 15, pp. 805–820, August 2009.

[6] R. Choudhury and N. Vaidya, "Deafness: a mac problem in ad hoc networks when using directional antennas," in *Proceedings of the IEEE ICNP Conference*, 2004.

[7] M. Takata, M. Bandai, and T. Watanabe, "RI-DMAC a receiver initiated directional mac protocol for deafness problem," *International Journal on Sensors Networks*, vol. 5, pp. 79–89, April 2009.

[8] J. Wang, H. Zhai, P. Li, Y. Fang, and D. Wu, "Directional medium access control for ad hoc networks," *Wireless Networks*, vol. 15, pp. 1059–1073, November 2009.

[9] A. Das and T. Zhu, "A reservation-based tdma mac protocol using directional antennas (RTDMA-DA) for wireless mesh networks," in *Proceedings of the IEEE GLOBECOM Conference*, 2007.

[10] J. Winters, "Smart antenna techniques and their application to wireless ad hoc networks," *IEEE Wireless Communications*, vol. 13, no. 4, pp. 77 –83, 2006.

 ISBN: 978-1-61208-126-7

On the Analysis of Packet Scheduling in Downlink 3GPP LTE System

Oana Iosif, Ion Banica
Faculty of Electronics, Telecommunications and Information Engineering
"Politehnica" University of Bucharest
Bucharest, Romania
oana_iosif@yahoo.com, banica@comm.pub.ro

***Abstract*—This paper investigates the performance of packet scheduling in downlink LTE (Long Term Evolution) systems using Round Robin strategy in time domain and time and frequency domain. Two types of non-real time services are considered in the analysis performed, with and without priority set, as well as the limitation given by the physical downlink control channels (PDCCH) on the number of simultaneously scheduled users. Cell throughput, achievable user throughput and system capacity are evaluated in different scenarios with two mixed services, two packet scheduling approaches and priority impact on the generated traffic.**

Keywords-LTE; OFDMA; PDCCH; packet scheduling; Round Robin

I. INTRODUCTION

Long Term Evolution (LTE) is the name given to a 3GPP (3rd Generation Partnership Project) concerning UTRAN (Universal Terrestrial Radio Access Network) evolution to meet the needs of future broadband cellular communications. This project can also be considered as a milestone towards 4G (Fourth Generation) standardization. The requirements set for LTE specified in [1] envisage high peak data rates, low latency, increased spectral efficiency, scalable bandwidth, flat all-IP network architecture, optimized performance for mobile speed, etc. In order to fulfill this extensive range of requirements several key technologies have been considered for LTE radio interface of which the most important are: multiple-access through Orthogonal Frequency Division Multiple Access (OFDMA) in downlink and Single Carrier - Frequency Division Multiple Access (SC-FDMA) in uplink and multiple-antenna technology.

Packet Scheduling is one of LTE Radio Resource Management (RRM) functions, responsible for allocating resources to the users and, when making the scheduling decisions, it may take into account the channel quality information from the user terminals (UE), the QoS (Quality of service) requirements, the buffer status, the interference situation, etc. [2]. Like in HSPA or WiMAX, the scheduling algorithm used is not specified in the standard and it is eNodeB (Evolved NodeB) vendor specific.

In this paper, we evaluate the performance of packet scheduling in downlink LTE using the Round Robin strategy through the results obtained for the average cell throughput, the achieved user throughput and the system capacity. These results may be considered in the LTE network design, in order to approximate the number of users that can be served with a certain throughput in a commercial LTE network.

The remainder of this paper is organized as follows. Section II discusses several aspects on scheduling and assigned resources in downlink LTE system followed by the Round Robin model description for different resource assignment approaches in Section III. Section IV depicts the results of the simulated scenarios and the conclusions are driven in Section V.

II. SEVERAL ASPECTS ON RESOURCE ALLOCATION IN DOWNLINK LTE

The benefit of deploying OFDMA technology on downlink LTE is the ability of allocating capacity on both time and frequency, allowing multiple users to be scheduled at a time. The minimum resource that can be assigned to a user consists of two Physical Resource Blocks (PRBs) and it is known as chunk or simply Resource Block (RB) [2],[3]. In downlink LTE one PRB is mapped on 12 subcarriers (180 kHz) and 7 OFDM symbols (0.5 ms) and this is true for non-MBSFN (Multimedia Broadcast multicast service Single Frequency Network) LTE systems and for normal Cyclic Prefix (CP). Scheduling decisions can be made each TTI (Time Transmission Interval) that in LTE is equal to 1 ms.

For non-real time services dynamic scheduling is usually used as it provides flexible and even full utilization of the resource.This scheduler performs scheduling decisions every TTI by allocating RBs to the users, as well as transmission parameters including modulation and coding scheme. The latter is referred to as link adaptation. The allocated RBs and the selected modulation and coding scheme are signaled to the scheduled users on the PDCCH (Physical Downlink Control Channel). The dynamic packet scheduler also interacts closely with the HARQ (Hybrid Automatic Repeat Request) manager as it is responsible for scheduling retransmissions and it may also take into account the QoS attributes and buffer information [4].

The channel conditions may or may not play a role in scheduling decisions. An alternative to channel-dependent scheduling is Round Robin strategy that serves the users in cyclic order, regardless the channel information.

Although OFDMA technology allows the users to be multiplexed in time and frequency, the scheduler, according to the implemented algorithm, may choose to allocate the entire bandwidth to a single user, reducing the scheduling to be done only in time domain. The channel-sensitive

 ISBN: 978-1-61208-126-7

scheduling done in time domain only is called Non-Frequency Selective Scheduling (NFSS) and the scheduling exploiting the channel variations in both time and frequency is known as Frequency Selective Scheduling (FSS) as specified in [5]. Fig. 1 illustrates an example of FSS for two users [6].

When scheduling is done in time and frequency domain, independently if it is channel-aware or not, the number of multiplexed users at each TTI is limited by the number of PDCCHs that can be configured. This depends on the system bandwidth, the number of symbols signaled for PDCCH allocation, the PDCCH format number, etc. [3], [4], [7], [8]. The PDCCHs are intended to provide both uplink and downlink scheduling information and as a PDCCH is allocated to each user to be scheduled, the maximum number of scheduled users per TTI in downlink LTE is half of the number of PDCCHs available. The authors from [4] discussed this constraint and proposed a three-step packet scheduling algorithm as it is depicted in Fig. 2.

III. ROUND ROBIN SCHEDULING MODEL IN DOWNLINK LTE

As mentioned in Section II, Round Robin scheduling is a non-aware scheduling scheme that lets users take turns in using the shared resources (time and/or RBs), without taking the instantaneous channel conditions into account. Therefore, it offers great fairness among the users in radio resource assignment, but degrades the system throughput. Both Time Domain Round Robin (TD RR) and Time and Frequency Domain Round Robin (FD RR) scheduling models are described in this section.

A. Time Domain Round Robin scheduling

In TD RR the first reached user is served with the whole frequency spectrum for a specific time period (1 TTI), not making use of the information on his channel quality and then these resources are revoked back and assigned to the next user for another time period. The previously served user is placed at the end of the waiting queue so it can be served with radio resources in the next round. This algorithm continues in the same manner [9]. Fig. 3 illustrates the resource sharing between two users with TD RR algorithm. The colors make the difference between the users. In this

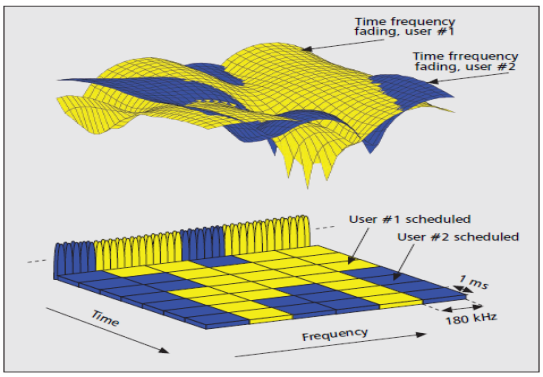


Figure 1. Frequency selective scheduling illustration for two users in downlink LTE

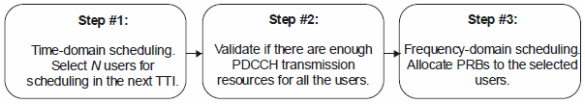


Figure 2. Illustration of a three step scheduling algorithm framework

example, every user is allocated 100% of the RBs and 50% of the time resource, so each gets 50% of the global resource.

Let us suppose a CBR (Constant Bit Rate) service of 250 kbps and a SNR (Signal to Noise Ratio) throughput per RB given by the radio conditions of 500 kbps. Assuming that there is one static user making the service and the same SNR is experienced in each RB and in all TTIs, the maximum amount of data that can be sent during one TTI per RB is 0.5 kb. Considering the system bandwidth of 20 MHz which consists of 100 RBs, the user needs to be allocated all resources for five TTIs to reach his service throughput. Therefore the user must be allocated 1/200 of the total resource in order to be served. This ratio is equal to service throughput / (SNR throughput*total number of RBs given by the system bandwidth). This represents the main idea in the TD RR model.

B. Time and Frequency Domain Round Robin scheduling

The FD RR allows multiple users to be scheduled within one TTI in cyclic order. Keeping in mind the PDCCH limitation discussed in Section II, the scheduling framework from Fig. 2 can be applied. The TDPS (Time Domain Packet Scheduling) may select N users in RR fashion to be scheduled in one TTI, but the PDCCH resources (M) must be checked in order to see if all users selected by the TDPS can be simultaneously scheduled. M users at most can be the input of FDPS (Frequency Domain Packet Scheduling), which schedules each user with RR strategy across different RBs. In the next TTI the users that were not selected in the previous one will be scheduled in the same manner and so on.

The FD RR is briefly presented in [10] where PDCCH constraint is not considered. The authors propose that all users be allocated one RB before reallocating to the same user. If the number of users waiting to be scheduled is less than the number of PDCCHs per TTI, this approach is correct. But if the users selected within one TTI are greater than the PDCCHs and if the idea of allocating one RB to each user is maintained, the result will be a waste of resources.

The resource sharing between two users with FD RR, assuming a hypothetical system bandwidth of two RBs, is depicted in Fig. 4. As in Fig. 3, each user is allocated 50% of the global resource. Taking the example given in Section III.A, but considering the limitation of 20 PDCCHs per TTI

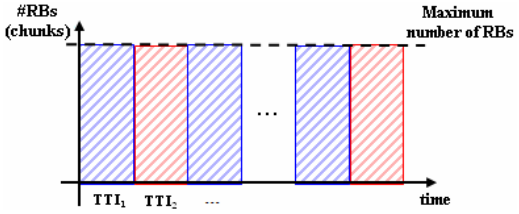


Figure 3. Resource sharing between two users with TD RR

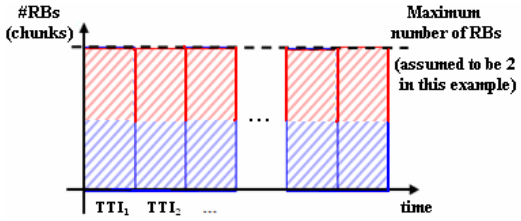


Figure 4. Resource sharing between two users with FD RR

for downlink LTE as it is concluded from [4], [7] and [8] and 40 users having the same radio conditions and making the same service, one user needs to be allocated 1 RB for 500 TTIs. The global resource in this case is reduced due to PDCCH constraint i.e. the maximum throughput given by the radio conditions * number of PDCCHs. The radio resource ratio assigned to each user is 1/40, higher than in TD RR example, so the capacity will be smaller. A solution to address this problem would be the allocation of more RBs at once to each user in order to exploit all transmission bandwidth. Knowing that for 20 MHz band in downlink LTE 20 users can be simultaneously scheduled at most, each user can be allocated 5 RBs before assigning resources to another one. In this case, the FD RR cell throughput will be the same as for TD RR, with the only advantage of being more suited to services with small packets and some delay requirements.

IV. SIMULATION SCENARIOS AND RESULTS

A computer simulation using C++ platform is conducted to evaluate the performance of RR scheduling in downlink LTE. For the simulations performed a single cell eNodeB is considered, with a carrier frequency of 2.6 GHz FDD (Frequency Division Duplex) and a system bandwidth of 20 MHz. The antenna configuration used in all scenarios is SISO (Single Input Single Output) which leads to category 1 terminal with ~10Mbps for all users. In order to reduce the complexity of the system simulations, we assume that equal downlink transmit power is allocated on each RB and all transmitted packets are received correctly. Moreover, the users are static and experience the same SNR values for all RBs allocated for all the simulation period (they are all located in the same bin).

The downlink SNR values used in this paper, resulting from the pathloss, shadow fading, multipath fading, eNodeB transmit power and thermal noise, are listed in Table I, along with the corresponding modulation and coding schemes and data rates.

The following sub-sections present the simulation results for cell throughput, average user throughput and system capacity in downlink LTE with RR scheduling. There are two categories of users considered: the first makes a CBR streaming service with 2 Mbps expected throughput (under this value the users cannot be served) and the second makes a VBR best effort service with 2 Mbps minimum accepted throughput, but it can reach more . The maximum best effort throughput reached is limited by the minimum between the data rate corresponding to the SNR experienced and the maximum throughput given by the user terminal category.

TABLE I. DOWNLINK SNR TO DATA RATE MAPPING

Minimum downlink SNR values (db)	Modulation and coding scheme	Data rate (kbps)
1.7	QPSK (1/2)	138
3.7	QPSK (2/3)	184
4.5	QPSK (3/4)	207
7.2	16 QAM (1/2)	276
9.5	16 QAM (2/3)	368
10.7	16 QAM (3/4)	414
14.8	64 QAM (2/3)	552
16.1	64 QAM (3/4)	621

A. Cell throughput results

Fig. 5 and Fig. 6 show the cell throughput with TD RR and FD RR for streaming users and best effort users.

The dependence of the cell throughput on the SNR values with 30 users in the cell is depicted in Fig. 5. An interesting evolution is shown by the cell throughput in FD RR for streaming service, where the cell saturation is reached. The explanation lies in both PDCCHs limitation of 20 per TTI and the CBR service of 2 Mbps. Despite the PDCCH limitation in FD RR for best effort users, cell saturation is not reached due to their capability of achieving a higher throughput compared to their service throughput. All 30 users are served only in TD RR for the last SNR throughput value.

When comparing TD RR with FD RR based on the results illustrated in Fig. 6 it can be concluded that for best effort users they show the same cell throughput evolution, This is not the case for streaming users because in TD RR the cell throughput is higher due to a higher number of users served. From the cell throughput saturation it can also be seen that in TD RR there are 31streaming users served, while in FD RR only 20 users reach their service requirements.

B. Average user throughput results

Fig. 7 shows the evolution of average user throughput with the number of users in the cell. For streaming service the user throughput is constant at 2 Mbps, while for best effort users it varies until the cell saturation is reached, the saturation point being the maximum number of users served. The maximum best effort user throughput is limited by the terminal category. The achievable best effort user throughput is higher in FD RR than in TD RR for more than 20 users in the cell because there are fewer users served and the cell resource is shared between a smaller number of users.

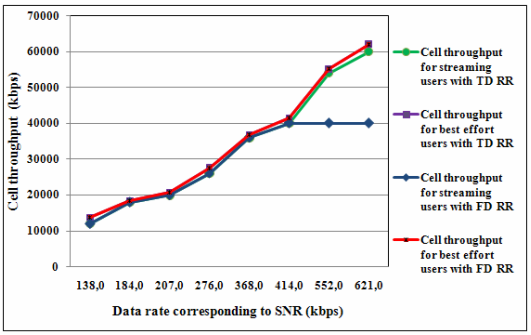


Figure 5. Cell throughput vs. SNR for TD RR and FD RR

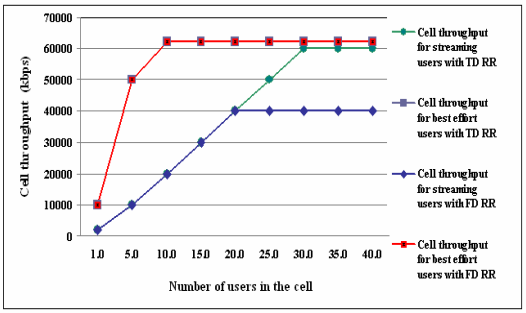


Figure 6. Cell throughput vs. the number of users in the cell for TD RR and FD RR

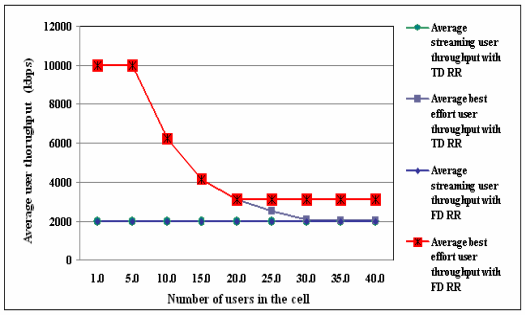


Figure 7. Average user throughput vs. the number of users in the cell for TD RR and FD RR

C. *System capacity results*

Fig. 8 and Fig. 9 show for both scheduling strategies how many users are served from the total number of users in the cell and the impact of the priority set for streaming service on the number and types of users scheduled. Half of the users in the cell are best effort users. The cell saturation is reached for 31 users served in TD RR and 20 in FD RR. When no priority is set, the number of served streaming users is equal to that of best effort users. For 50 users in the cell and priority set, in TD RR there are 6 best effort users and 25 streaming users served, while in FD RR there is no best effort user served and 20 streaming users served.

V. CONCLUSIONS AND FUTURE WORK

This paper illustrates the performance of Time Domain Round Robin and Time and Frequency Domain Round Robin scheduling in downlink LTE in what concerns the cell throughput, the average user throughput and the system capacity, in the assumption of two types of services with the possibility of applying a higher priority to one of them, having an impact on which number and type of users are served. The constraint of PDCCHs on the number of users scheduled each TTI has also been outlined and depicted in the simulation results, making FD RR less efficient when the number of users in the cell is higher than the PDCCHs. These results may be considered when designing a real LTE network. Future work will focus on the analysis of packet scheduling in uplink LTE system.

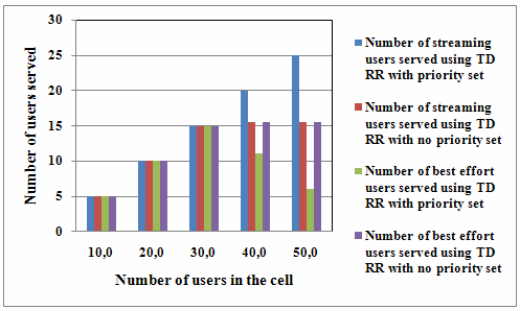


Figure 8. Number of users served vs. number of users in the cell w/o priority in TD RR

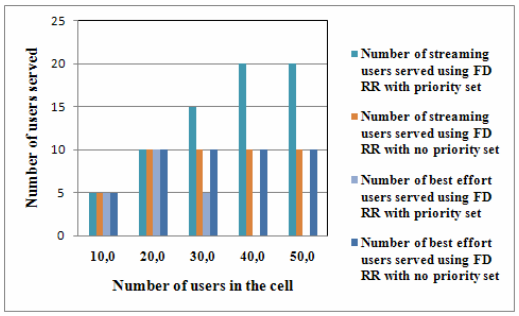


Figure 9. Number of users served vs. number of users in the cell w/o priority in FD RR

ACKNOWLEDGMENT

Oana Iosif is POSDRU grant beneficiary offered through POSDRU/6/1.5/S/16 contract.

REFERENCES

[1] 3GPP TR 25.913v8.0.0 Release 8,"Requirements for evolved UTRA (E-UTRA) and evolved UTRAN (E-UTRAN)".

[2] 3GPP TS 36.300v8.12.0 Release 8, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); Overall description; Stage 2".

[3] 3GPP TS 36.211v8.9.0 Release 8,"Evolved universal terrestrial radio access (E-UTRA); Physical channels and modulation".

[4] H. Holma and A.Toskala, "LTE for UMTS: OFDMA and SC-FDMA based radio access", 2009 John Wiley & Sons, pp.181-212.

[5] F. Khan, "LTE for 4G mobile broadband", Cambridge University Press 2009.

[6] D. Astély, E. Dahlman, A. Furuskär, Y. Jading, M. Lindström, and S. Parkvall, "LTE: The evolution of mobile broadband", Communications Magazine, IEEE In Communications Magazine, IEEE, Vol. 47, No. 4. (05 May 2009), pp. 44-51.

[7] R.Love, R. Kuchibhotla, A. Ghosh, R.Ratasuk, B. Classon, Y. Blankenship, "Downlink control channel design for 3GPP LTE", Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE, pp. 813-818.

[8] D. Laselva, F. Capozzi, F. Frederiksen, K. I. Pedersen, J. Wigard and I.Z. Kovács, "On the impact of realistic control channel constraints on QoS provisioning in UTRAN LTE", Vehicular Technology Conference Fall, 2009 IEEE 70th, pp. 1-5.

[9] S. Hussain, "Dynamic radio resource management in 3GPP LTE", Blekinge Institute of Technology 2009.

[10] C. Han, K. C. Beh, M. Nicolaou, S. Armour, A. Doufexi, "Power efficient dynamic resource scheduling algorithms for LTE", Vehicular Technology Conference Fall, 2010 IEEE 72nd, pages 1-5.

 ISBN: 978-1-61208-126-7

Accelerating Cryptographic Protocols: A Review of Theory and Technologies

Antti Hakkala[1,2] and Seppo Virtanen[2]

Turku Centre for Computer Science TUCS[1]
Department of Information Technology[2]
University of Turku
Turku, Finland
Email: {antti.hakkala,seppo.virtanen}@utu.fi

***Abstract*—Modern cryptography applications require significant processing power and resources on computers. To make implementations of these algorithms comply with the rising requirements of speed and throughput of modern applications, the use of instruction set extensions and external cryptography processors has become more and more commonplace. While cryptography algorithms can and do differ significantly in their design and functionality, there are certain algebraic operations which are used throughout all algorithm types. In this paper, we review the fundamental operations that can be found behind several cryptography algorithms currently in use. We examine the need for enhancing the performance of cryptographic protocols and available methods for accelerating the computation of such algorithms. We discuss current methods for accelerating their performance and examine the use of instruction set extensions for cryptography algorithms, particularly the cases where an instruction set can be used for multiple purposes. We conclude that future applications require making these instruction sets as general as possible to support a wide range of algorithms.**

Keywords-Cryptography; Instruction Set Extensions; Embedded Security; ASIP.

I. INTRODUCTION

Cryptography is the science—and some say, art [1]—of hiding and securing information. Cryptography has an interesting and far-reaching history dating back thousands of years, but what can be considered modern cryptography began in the 20th century. Driving forces to this development were the advent of computers and the pressure from the second World War, where cryptography and cryptanalysis played a vital role in its' outcome [2][3]. Currently, the whole global communication network is heavily dependent on cryptography and its applications.

As a result of this progress, applications handling personal data, banking information, spatial data and private communications have spread to mobile and embedded devices. This data must be secured at all points in the communication chain. Cryptography is computationally intensive, and thus modern embedded devices that must be able to process encrypted information require different cryptography acceleration methods to function properly in this environment. There are many different possible approaches to solving this problem, and in this paper we consider the problem with focus on Instruction Set Extensions (ISE) for a general-purpose CPU (GPCPU). This approach allows the device to execute certain pre-defined operations more effectively than a software approach. A dedicated hardware solution on silicon would be faster, but these solutions sacrifice flexibility and programmability for speed. Properly designed ISEs can provide concrete improvements to performance with additional programmability added to the system, but before this is possible, the central building blocks from number theory that are required for cryptography algorithms must be identified. If these blocks can be mapped to other, seemingly unrelated functions, the instruction sets can be designed to accommodate even broader functions than cryptography.

In this paper we review and examine the building blocks of cryptography algorithms at different security model abstraction levels. We discuss current methods for accelerating their performance and examine the use of instruction set extensions for cryptography algorithms. We start the discussion by defining the layered security model we use for categorizing security elements in this paper in Section II. In Section III, we proceed to give an introduction to the foundations of ISEs. Then, in Section IV, we proceed to review the predominant types and functions of modern cryptographic algorithms. In Section V, we review and discuss methods for accelerating the processing of cryptographic protocols by examining the mathematical foundations behind cryptographic algorithms. In Section VI, we review proposed solutions for cryptographic instruction set extensions, including co-processor and application-specific instruction-set processor (ASIP) approaches. In Section VII, we suggest some of the interesting future research directions regarding embedded security, cryptographical primitives and ISE design for embedded systems. Our concluding remarks are discussed in Section VIII.

II. THE PYRAMID MODEL OF SECURITY

The structure of any information security design can be generalized into the form of a security pyramid [4], divided into five different levels ordered by the level of abstraction, from highest to lowest. At the top abstraction level is the security protocol architecture, which describes the protocols used for secure communications. In the next layer are the different algorithms used to achieve the security specified by

 ISBN: 978-1-61208-126-7

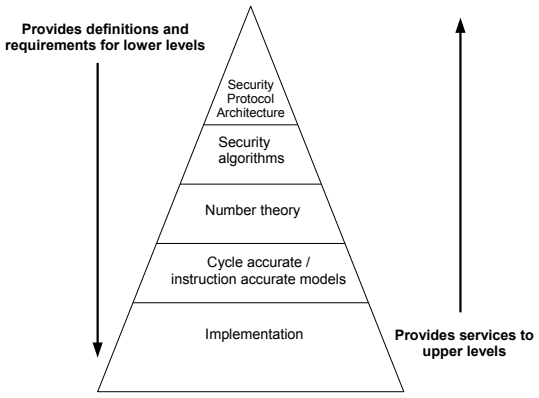


Fig. 1. Security Pyramid [4]

the protocols. These algorithms include the basic signature, encryption and hash algorithms required for cryptography. The third level contains the basic building blocks of the algorithms themselves. These include the number theoretic foundations of the algorithms, and different number representations for speeding up cryptography algorithm implementations. The fourth level contains the platform-independent representations of the algorithms, and the fifth and final level is the actual physical silicon implementation of the algorithms in the protocol architecture.

The security pyramid model can also be seen as a service architecture. The upper levels define requirements (what needs to be done) and general architecture guidelines (how this can be done), and the lower levels provide services upwards by implementing the requirements of the upper level. For example, when the security protocol is chosen, it defines what security algorithms will be required for implementation, but it does not explicitly specify *how* they should be implemented, as long as they meet the requirements. Similarily, the selection of security algorithms leave the designer free to explore actual algorithm implementation options based on the requirements from the upper level (optimization for speed, energy efficiency, side-channel resistance). The instruction accurate level provides the actual number theoretic functionalities of the third level in the form of instruction sets. Finally this continues down to the implementation level, where the requirements and specifications from the upper levels define the actual physical implementation and layout of the device.

The interactions between levels are straightforward when the security pyramid is considered to be separate from the rest of the design process of the whole device the security features are a part of. When the whole process is considered, this model is not accurate enough to represent the problems and relations in security design for devices.

A similar security pyramid for embedded systems has been presented in [5]. It also contains five levels of abstraction, but with little differences. The top level, similarly to the previous model, is the protocol level. The second level is the algorithm level. The third level is the architecture level, which consists of secure design architectures and partitioning to prevent attacks against the systems from software. The fourth level is the microarchitecture level, which contains secure hardware design guidelined by the architecture level decisions. The lowest level is the circuit level, which contains secure physical layer implementations for the chosen microarchitecture. In this model, number theory has been absorbed by the algorithm level to make way to secure design and partitioning, an important issue with secure embedded systems.

What is common for both models is that, as is pointed out in [5], the whole problem of secure system design is not necessarily concentrated on a single level of abstraction. Some problems can be addressed on a single level, but others must be addressed on several levels or the security of the system breaks down. Because adversaries usually have full physical access to an embedded device they are trying to compromise, the security must be sound on all levels on the security pyramid, or the whole system is compromised. This is challenging from a design standpoint, and more research needs to be directed to ensuring built-in system security in all abstraction layers and all phases of the design process.

III. Introduction to Instruction Set Extensions

Cryptographic algorithms utilize certain algebraic operations, which are often not directly implemented in or supported by the *Instruction Set Architecture* (ISA) of modern digital systems. Cryptography applications require a large amount of computation resources, so identifying and streamlining the calculation of these algebraic operations is essential to good performance on digital computers. This can be implemented with software, but this approach is slow and requires a lot of power. Conversely, pure hardware solutions are very fast and efficient, but lack the flexibility of software. Thus hybrid methods that use both techniques are required. There are several methods available to achieve performance gains, and they are dependent on the overall architecture and purpose of the target system [6]. General purpose embedded processors can be paired up with co-processors, FPGA solutions or hardware accelerators, or specifically designed ASICs or ASIPs can handle the extra computational load. This review is centered on specific processor ISEs used to handle these operations. ISEs expand the available native instructions for a processor, and when an operation is supported by a specific ISE, its' processing is substantially faster than an implementation purely based on software.

ISEs can be mapped to the security pyramid, where they occupy the fourth level, as is seen on Figure I. As ISEs are dependent on the building blocks provided by their respective upper layers, the fundamentals of these operations must be understood so that the best possible choices can be made for the basis of ISE design. These are naturally dependent on other architectural choices, so this review will take into account the most basic and general methods.

When considering ISEs and their implementation, it must be realized that the pressure for effective implementation comes from the upper levels of the security pyramid. At the implementation level, there is not much we can do after

 ISBN: 978-1-61208-126-7

the specifications have been decided and the chip design is finished. Similarly, there is nothing a designer can do about cryptography protocol or specific algorithm design, the designer must work between these layers to provide good service to both upper and lower layers in the security pyramid.

An instruction for a single round of a block cipher can be implemented, and it provides very fast performance for that single block cipher. What happens if the cipher suite needs to be changed by altered application demands? Even though a new device can be brought in to handle the changed requirements, it is not always practical to change the hardware in applications for which the cipher suites can change regarding to the service they provide. If the device is required to handle several different cryptography protocols for different purposes, it is desirable to have as broad a functionality as possible in the single chip without losing performance significantly. Also, the designer must pay attention to related functions that overlap with the security features [4]. This means that the instructions provided for cryptography applications can be used for other purposes as well, if the underlying functions have sufficient similarities, such as using the same permutations or field operations.

IV. Modern Cryptography Algorithms

Contemporary communication systems are heavily dependent on secure communication of information. Algorithms for secure communications have been developed throughout history, but many of them have been broken by cryptanalysis. Here we focus on the three main groups of cryptography algorithms. When considering an algorithm for cryptography purposes, a designer has a wide range of choices available, as the number of available and secure ciphers is very large and growing constantly. Cryptographic algorithms can be broadly divided into *symmetric* or *private-key*, *asymmetric* or *public-key* ciphers and *cryptographic hash functions*.

The two basic techniques for obfuscating the relationship between plaintext and ciphertext are *confusion* and *diffusion* [7]. Confusion aims to blur the relationship between the plaintext, key and ciphertext, while diffusion eliminates the statistical redundancy inherent in the plaintext and distributes it over the ciphertext. Good ciphers should employ both techniques, and several different operations for both public and private key systems exist for achieving confusion and diffusion. Symmetric cryptography algorithms use the same key for encryption and decryption. Before two communicating parties can begin to exchange encrypted messages, they have to agree on the algorithm to be used and a common key both will use for encryption and decryption. After a key has been agreed upon, both parties use the same key for both encryption and decryption.

In 1976, Whitfield Diffie and Martin Hellman introduced the concept of public-key cryptography [8]. A public key algorithm has a pair of keys, one for encryption (public key), the other for decryption (private key). The public key for encrypting messages to the recipient is published and the private key for decrypting received messages is kept secret. The security of public-key cryptography lies on a difficult mathematical problem that is easy to solve with certain specific information, but next to impossible otherwise. The intractability of *integer factorialization* and *discrete logarithms* in certain groups are the most used problems as the basis of public-key cryptography algorithms. Some problems thought at first to be intractable have been shown to be solvable, thus breaking the cryptosystem built on it.

Hash functions are essential to many cryptographic protocols, because they provide condensed versions of their input in a manner which cannot be reversed. This has several applications in message authentication schemes and integrity checks, among other important functions. An ideal hash function is a one-way function that generates a fixed length *digest* $h(m)$ of an arbitrary length input value m. The one-wayness of a function means that evaluating the function is easy, but evaluating its' inverse is not feasible.

V. Methods and Algorithms for Accelerating Cryptographic Protocols

To effectively realize fast implementations for cryptography algorithms, it is necessary to examine the number theoretical foundations of these algorithms. The following approaches are mainly algorithmical or mathematical approaches to accelerating the calculation of central operations in cryptography. In addition to the fundamental basis of increasing speed of cryptographic calculations, the understanding of these operations also help to give insights into cipher design and requirements.

The techniques examined here address modular arithmetic, point scalar multiplication and operations for symmetric ciphers. They are based on the third level of the security pyramid, and form the basis of instruction set extensions found on the fourth level. As such, they essentially define how fast we are able to perform the required operations for cryptography algorithms in the first place.

A. Exponentiation and modular arithmetic

Exponentiation is a central arithmetic operation in most algorithms, and consequently the speed of exponentiation significantly affects their performance. Modular arithmetic operates in a residual number system, where all operands are reduced *modulo* an integer. Cryptography combines these two central operations in many ciphers, and because of this, they are critical to the performance of cryptography applications.

Trivial multiplication algorithms have a complexity of $O(n^2)$ [9]. Algorithms such as Karatsuba-Ofman multiplication [10] have a complexity of $O(n^{lg\ 3})$. Efficient exponentiation has many solutions, and a good review of fast methods and their mathematical fundamentals can be found in [11], where addition chains, the binary method, m-ary method, different window methods and redundant number systems are considered. It was found in the review that the choice of the multiplication algorithm depends on the situation. Sometimes, as is with Diffie-Hellman key exchange, it is more prudent to use precomputation methods, while in other situations the choice of the group will make all the difference, as is with $\mathbb{F}_{2^n}$

 ISBN: 978-1-61208-126-7

in Elliptic Curve Cryptography (ECC), where it is possible to use different coordinate systems for a substantial advantage. Other number representations for exponentiation have been proposed, such as the Fibonacci m-ary representation [12], which is a sparse number representation based on representing numbers as sums of Fibonacci numbers.

The Chinese Remainder Theorem (CRT) is often used to enhance RSA performance, as it can be used to effectively halve the required modulus length by calculating exponentiations $\mod p$ and $\mod q$ instead of $\mod (pq)$. A description of CRT can be found in, for example, [1]. All methods where the modulus is a composite and require modular exponentiation can benefit from using CRT, with some reservations. For example RSA decryption greatly benefits from this, but encryption does not, as the factors of the composite modulus $n = pq$ are required for the CRT, and the security of RSA depends on only the receiver knowing the factors for n. The effect of using CRT on RSA results in theoretical decryption speed increase of 4x, and for the implementation in [13], actual speedup is 3.8x.

The use of modular arithmetic requires modular reduction and inversion operations. These are computationally expensive, and methods to improve their performance are well-documented. Montgomery presented a method [14] for accelerating modular exponentiation by transforming the calculations into a new representation where divisions and reductions by n can be performed as divisions and reductions by some arbitrary binary number 2^i. Given that reductions and divisions by powers of 2 are trivial in digital computers, this greatly increases the speed of modular exponentiation. Although the benefit is lost for minor operations, as the cost of transforming the operands to the Montgomery representation is not free, repeated operations benefit from this approach greatly. Discussion and algorithms for Montgomery reduction and multiplication can be found in, for example, [15]. The Montgomery method is widely implemented in digital systems of all kind where efficient modular reduction is required, and improvements to the original algorithm are widely researched. An algorithm presented in [16] runs faster than Montgomery's, and differs from the Montgomery method by integrating the multiplication and reduction steps. In [17], a modified version of Montgomery multiplication is presented, with a reported reduction in multiplication steps of 26.68% to 38.9% compared to previous improvements.

An alternative method for fast modular reduction [18] was presented by Barrett. It is based on precomputing a single parameter for later use. This precomputation cost can be considered negligible if the modulus does not change during calcualtions. So far, the Montgomery method is more prominent in existing implementations. A modified Barrett implementation was presented in [19]. It is based on modifying the precomputation phase and by implementing a modified Karatsuba-Ofman multiplication algorithm to provide substantial improvement to performance compared to the Montgomery method. When compared on three different platforms, the performance of the modified Barrett algorithm was consistently better than Montgomery's, and increased with larger moduli.

Calculating multiplicative inverses are an important part in any modular arithmetic based cryptosystem. This can be done via Fermat's Little Theorem, which states that $a^{p-1} \equiv 1 \mod p$ and thus the multiplicative inverse of element a is $a^{p-2} = a^{-1} \mod p$. An another way to calculate inverses is the Extended Eucleidian Algorithm, of which a description can be found in, for example, [20].

B. Elliptic Curve Cryptography

Elliptic Curve cryptosystems are based on elliptic curves over finite fields [21], [22]. A basic introduction to ECC can be found in [23], and a more comprehensive treatment in [20]. While cryptosystems operating in $\mathbb{Z}_p$ use modular exponentiation, the central operation for ECC is point scalar multiplication, where a point P on the elliptic curve is added to itself k times. The reversal of this calculation is known as the Elliptic Curve Discrete Logarithm Problem (ECDLP), and the intractability of this problem in the chosen underlying field is the basis of ECC. Attacks on ECC exist—like RSA, the ECDLP has not been *proven* to be intractable—but they can be alleviated with proper choice of system parameters [20].

Modular arithmetic is present in ECC systems, as the operands are reduced modulo a prime for prime fields $\mathbb{F}_q$, and modulo an irreducible polynomial in extension fields $\mathbb{F}_{2^n}$. Similar m-ary techniques, windowing techniques and NAF representations discussed in Section V-A can be used for accelerating such calculations [23]. The coordinate system used to represent the points on an elliptic curve can also be modified to provide faster calculations by reducing or removing the need for inversions, which are costly to compute. Itoh-Tsujii inversion [24], derived from Fermat's Little Theorem, can be used to calculate inversions in both prime and extension fields. The use of repeated squarings to compute inversions in binary fields is discussed in [25]. Also, the choice of both the curve and the underlying field can have a significant effect in speeding up calculations. See [20] for discussion on accelerating ECC systems.

VI. Instruction Set Extensions for Cryptography

Given the demonstrated complexity of operations in cryptography algorithms and the fact that computers are used by necessity to perform these operations, the optimization of performance for computer systems is essential for any cryptography application. Wu et al. [26] state that for a fast and efficient platform for cryptography applications, the platform must be scalable and it must perform the essential operations required in cryptography efficiently.

ASIPs have been intensively studied already for more than a decade. The aim in ASIP design is to find sequences of general purpose operations in the target application, and group these sequences into a hardware implementation [27], [28], [29], [30]. Often the target application is profiled using tool-based automation to find such code sequences. The resulting recurring command sequences that are chosen for hardware implementation are often quite short, usually less than 10

 ISBN: 978-1-61208-126-7

and often only 2-3 general purpose processor commands of length. In ASIP methodologies often a general purpose processing core is enhanced with hardware execution units for the detected command sequences with the aim that the special hardware units speed up overall application processing speed considerably. For example, in an early study presented in [27] it was determined that this kind of an approach provides a performance increase of no more than 30 % when compared to a general purpose processor designed with equal area and power constraints.

ASIP design methodologies have established their position as an attractive alternative to embedded processor design, especially in the digital signal processing application domain. The use of ASIPs for cryptographic applications has also been studied quite intensively in recent years. For example, in [31], an ASIP for speeding up AES, DES, MD5 and SHA processing is designed. The design process attempts to identify all possible levels of parallelism in the target application both at the instruction level (where detected sequences are 5-10 instructions of length) and at the task level, thus aiming for a better result than in the traditional instruction level profiling. The results showed up to 2.7x cycle time improvement for the target algorithms. In [32], an existing ASIP design platform is used to design an ASIP for AES processing in sensor networks. The optimization includes the design of two new hardware units to speed up the AES MixColumns and SubBytes transformations. In comparison to an unoptimized reference model, the resulting optimized ASIP and its code show cycle count reductions of 33 % and 45 % for encryption and decryption, respectively. In [33], the design space is explored to design an ASIP for cryptographic pairings based on an existing RISC core. Based on the exploration, the key extensions to the core are a multi-cycle scalable Montgomery multiplier and an enhanced memory architecture. The resulting ASIP instances range from a fast ASIP instance suitable to embedded systems to a smaller and slower instance suitable to for example smart card applications.

Instruction set design can also affect the choice of underlying algorithms and methods dramatically. In [34], an instruction set is customized to provide superior multiplication performance with an algorithm clearly inferior to other available options when considered without the instruction set.

The question whether to accelerate public or private key algorithms, or both, is heavily dependent on the target application. For example the effects on cryptography acceleration on SSL is analyzed in [35], where it was found that given very short SSL sessions, public-key traffic for key exchange and setup dominates the total amount of encrypted traffic. In SSL, public-key encryption is used for key exchange and authentication, and symmetric encryption is used for payload encryption. As the amount of traffic increases, symmetric encryption becomes increasingly dominant in the overall traffic. If a server has to handle multiple distinct encrypted connections for a prolonged amount of time, public-key encryption traffic becomes increasingly important. For a client which has to handle only one connection, symmetric encryption becomes more and more meaningful as the amount of traffic grows.

A. Synergy benefits from instruction sets

In [36], a method for accelerating AES with an instruction set designed for ECC is presented. It exploits the fact that AES uses finite field arithmetic in the field GF(2^8), and thus ECC accelerators for fields of the form GF(2^m) can be used also for accelerating AES. The instruction set contained an instruction for binary polynomial multiplication, and this was used to enhance performance. The authors report a speed increase of up to 25% for AES when utilizing instructions for ECC in a SPARC V8 LEON-2 processor. Although the speed increase is not on par with dedicated AES implementations on FPGA:s, for example, it shows how it is possible to gain performance from exploiting the common building blocks of cryptographic algorithms.

The effects of the Intel AES instruction set [37][38] for performance on SHA-3 hash function candidates is analyzed in [39]. A significant majority of the SHA-3 candidates are directly based on or utilize the AES round operation, so direct performance improvements on the operation of such functions is intuitive. The analysis is based on simulations on the information available of the Intel AES ISE. As the instructions perform one complete round of AES, algorithms which do not directly comply to the round function do not gain performance benefits from these instructions, even though they have similar building blocks. This highlights the disadvantages of very specialized instruction sets; it is always a tradeoff between flexibility and performance, as the instructions perform a complete round, any deviation from the standard round causes the instructions to be unusable, even though the underlying operations are the same.

It would be reasonable to assume that any methods for accelerating AES performance would also have similar impact on the performance of the SHA-3 candidates that are based on AES. Especially instructions for ECC are interesting in this regard, as it has been previously noted that they can be used to accelerate AES, and thus would be by extension capable of doing the same to some of the SHA-3 candidates. To the best of our knowledge, no research in this direction has yet been published, though.

VII. Open Issues and Future Work

Because the capability of devices to execute a wide variety of cryptographic protocols effectively is directly tied to the security and privacy aspects of future communication systems, the creative use of instruction sets to enhance the performance of cryptography operations is an interesting research topic. This is especially true for devices which must provide a wide variety of functions using limited processing capabilities.

When considering the larger scale of secure system design, overcoming the information security challenges of embedded communication systems of the future requires continuous fundamental and cross-disciplinary research. As an enabling factor to approaching these challenges we see the need for a paradigm shift: instead of treating security as an incremental

 ISBN: 978-1-61208-126-7

feature of embedded communication systems, research efforts should focus on pre-empting the emergence of information security threats and vulnerabilities in future embedded communication systems. This requires incorporating information security as an integral part into the design process of such systems and their subsystems. Such a paradigm shift raises challenging novel research problems in both hardware and software development as well as related processes. Focusing research efforts on security-enabled hardware/software codesign solutions and systematic approaches for designing secure embedded communication systems would be an important first step in this direction.

In our view, the initial step towards more robust and secure system design could be integration of the design-time "building security in" philosophy [40] from software engineering to the run-time ability of adapting to new application-domain specific applications, as conceptualized in Software-Defined Radio (SDR) [41] and Open Wireless Architecture (OWA) [42], in the overall embedded communication system design process [43]. Bringing these research directions together to design secure embedded communication systems would be an extremely interesting approach. Some aspects of the research required to accomplish this are outlined in [44]. A thorough overview of design challenges for future embedded communication systems including their security is given in [45].

Security interacts directly with the price, power consumption, testability, performance and reliability factors of the design process [23]. This poses steep requirements to system design processes and methodologies. To achieve these goals, the requirements for such design processes should be explicitly defined and examined. Building security into embedded systems has effects in all vital areas of system design, and this is an interesting direction for future research.

VIII. Conclusion

Effective methods for accelerating different cryptosystems have been presented in literature. It remains as the designers dilemma to select the best of these methods which are applicable to the current problem and use them effectively to accelerate the required operations used in the design. The design process of an embedded system should incorporate security features to the core of the process, thus moving towards design methods where security is already built in. Security models such as the security pyramid are efficient in pointing out that security is not a feature that can be added in as an afterthought, but is a complex problem that involves all layers of abstraction.

Implementing efficient methods for essential algebraic and other operations can make a large difference in cryptography application performance. The choice of these operations depend on the nature of the cryptography application, but as some operations are central to many different algorithms, synergy benefits from implementing a broad base of operations are worth consideration for implementations where flexibility is a design parameter.

Instruction set extensions for general purpose processors have been demonstrated to be very efficient in improving cryptography application performance by providing instructions for particular operations that are ill-suited for execution in traditional CPUs. Performance increases have been studied and have been observed to be even an order of magnitude faster than implementations in pure software.

Instruction sets for particular algorithms are naturally faster in performance than generalized approaches, but fixed word lengths and logic cause loss of flexibility in choosing different algorithms for implementation. Instruction sets that target some general problematic operation that is shared between different ciphers are naturally more flexible, while still providing significant performance increase compared to pure software solutions. If the goal is to create a flexible solution for cryptography acceleration, these generalized instruction sets are a very tempting choice over specific instruction sets.

References

[1] B. Schneier. *Applied Cryptography: Protocols, Algorithms and Source code in C, 2nd edition*. John Wiley & Sons, 1996.

[2] D. Kahn. *The Codebreakers*. Macmillan Publishing Company, New York, 1967.

[3] S. Singh. *The Code Book: The Evolution of Secrecy from Mary, Queen of Scots, to Quantum Cryptography, 1st edition*. Doubleday, New York, NY, USA, 1999.

[4] P. Schaumont and I. Verbauwhede. A reconfiguration hierarchy for elliptic curve cryptography. In *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, 2001.*, volume 1, pages 449–453. IEEE, 2002.

[5] D.D. Hwang, P. Schaumont, K. Tiri, and I. Verbauwhede. Securing embedded systems. *Security and Privacy, IEEE*, 4(2):40 –49, March-April 2006.

[6] P. Kocher, R. Lee, G. McGraw, and A. Raghunathan. Security as a new dimension in embedded system design. In *Proceedings of the 41st Annual Design Automation Conference, San Diego, CA, USA, June 07 - 11*. ACM, New York, 2004.

[7] C.E. Shannon. Communication theory of secrecy systems. *Bell Syst. Tech. J.*, 28:656–715, Oct. 1949.

[8] W. Diffie and M. Hellman. New directions in cryptography. *IEEE Transactions on Information Theory*, IT-22:644–654, 1976.

[9] D.E. Knuth. *The art of computer programming. Volume 2. Seminumerical algorithms, Third Edition*. Addison-Wesley Reading, MA, 1998.

[10] A. Karatsuba and Y. Ofman. Multiplication of multidigit numbers on automata. In *Soviet Physics Doklady*, volume 7, page 595, 1963.

[11] D.M. Gordon. A survey of fast exponentiation methods. *Journal of Algorithms*, 27(1):129 – 146, 1998.

[12] S.T. Klein. Should one always use repeated squaring for modular exponentiation? *Information Processing Letters*, 106(6):232–237, 2008.

[13] K. Hansen, T. Larsen, and K. Olsen. On the Efficiency of Fast RSA Variants in Modern Mobile Phones. *International Journal of Computer Science and Information Security*, 6(3):136–140, 2009.

[14] P.L. Montgomery. Modular Multiplication Without Trial Division. *Mathematics of Computation*, 44(170):519–521, 1985.

[15] A.J. Menezes, P.C. Van Oorschot, and S.A. Vanstone. *Handbook of applied cryptography*. CRC, 1997.

[16] S.M. Hong, S.Y. Oh, and H. Yoon. New modular multiplication algorithms for fast modular exponentiation. In *Advances in Cryptology—EUROCRYPT96*, pages 166–177. Springer, 1996.

[17] C.-L. Wu. An efficient common-multiplicand-multiplication method to the montgomery algorithm for speeding up exponentiation. *Information Sciences*, 179(4):410 – 421, 2009.

[18] P. Barrett. Implementing the Rivest Shamir and Adleman public key encryption algorithm on a standard digital signal processor. In *Advances in Cryptology—CRYPTO86*, pages 311–323. Springer, 1987.

[19] W. Hasenplaugh, G. Gaubatz, and V. Gopal. Fast modular reduction. *18th IEEE Symposium on Computer Arithmetic (ARITH'07)*, 2007.

 ISBN: 978-1-61208-126-7

[20] D.R. Hankerson, S.A. Vanstone, and A.J. Menezes. *Guide to elliptic curve cryptography*. Springer-Verlag New York Inc., 2004.
[21] V. Miller. Use of elliptic curves in cryptography. *Proceedings of the Advances in Cryptology—CRYPTO'85*, 1986.
[22] N. Koblitz. Elliptic curve cryptosystems. *Mathematics of Computation*, 48:203–209, 1987.
[23] C.H. Gebotys. *Security in Embedded Devices*. Springer Verlag, 2010.
[24] T. Itoh and S. Tsujii. A fast algorithm for computing multiplicative inverses in GF(2^m) using normal bases. *Information and Computation*, 78(3):171–177, 1988.
[25] K.U. Järvinen. On repeated squarings in binary fields. In *Proceedings of the 16th International Workshop on Selected Areas in Cryptography, SAC 2009*, volume 5867 of *Lecture Notes in Computer Science*, pages 331–349. Springer-Verlag, 2009.
[26] L. Wu, C. Weaver, and T. Austin. CryptoManiac: a fast flexible architecture for secure communication. *ACM SIGARCH Computer Architecture News*, 29(2):110–119, 2001.
[27] M. Arnold and H. Corporaal. Designing domain specific processors. In *Proceedings of the 9th International Symposium on Hardware/Software Codesign (CODES'01)*, pages 61–66, 2001.
[28] A. Both, B. Biermann, R. Lerch, Y. Manoli, and K. Sievert. Hardware-software-codesign of application specific microcontrollers with the ASM environment. In *Proceedings of the Conference on European Design Automation*, pages 72–76, Grenoble, France, September 1994.
[29] T.V.K. Gupta, P. Sharma, M. Balakrishnan, and S. Malik. Processor evaluation in an embedded systems design environment. In *Proceedings of the 13th International Conferenence on VLSI Design*, pages 98–103, Calcutta, India, January 2000.
[30] J. Van Praet, G. Goossens, D. Lanneer, and H. De Man. Instruction set definition and instruction selection for ASIP. In *Proceedings of the Seventh International Symposium on High-Level Synthesis*, pages 11–16, Niagara-on-the-lake, Canada, May 1994.
[31] D. Montgomery and A. Akoglu. Methodology and Toolset for ASIP Design and Development Targeting Cryptography-Based Applications. In *Proceedings of IEEE International Conference on Application-Specific Systems, Architectures and Processors*, pages 365–370, Montreal, Canada, July 2007.
[32] N. Suarez, G.M. Callico, R. Sarmiento, O. Santana, and A.A. Abbo. Processor customization for software implementation of the AES algorithm for wireless sensor networks. In *Proceedings of the 19th International Workshop on Power and Timing Modeling, Optimization and Simulation, LNCS 5953*, pages 326–335, Delft, The Netherlands, September 2009.
[33] D. Kammler, D. Zhang, P. Schwabe, H. Scharwaechter, M. Langenberg, D. Auras, G. Ascheid, and R. Mathar. Designing an ASIP for Cryptographic Pairings over Barreto-Naehrig Curves. In *Proceedings of 11th International Workshop on Cryptographic Hardware and Embedded Systems, LNCS 5747*, pages 254–271, 2009.
[34] J. Großschädl, P. Ienne, L. Pozzi, S. Tillich, and A.K. Verma. Combining algorithm exploration with instruction set design: a case study in elliptic curve cryptography. In *Design, Automation and Test in Europe, 2006. DATE'06. Proceedings*, volume 1, pages 1–6. IEEE, 2006.
[35] J. Burke, J. McDonald, and T. Austin. Architectural support for fast symmetric-key cryptography. In *Proceedings of the ninth international conference on Architectural support for programming languages and operating systems*, pages 178–189. ACM, 2000.
[36] S. Tillich and J. Großschädl. Accelerating AES using instruction set extensions for elliptic curve cryptography. *Computational Science and Its Applications–ICCSA 2005*, pages 665–675, 2005.
[37] S. Gueron. Intel® Advanced Encryption Standard (AES) Instructions Set. White paper, Intel Corporation, http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-aes-instructions-set/, January 2010.
[38] S. Gueron. Intel's New AES Instructions for Enhanced Performance and Security. In *Fast Software Encryption: 16th International Workshop, FSE 2009 Leuven, Belgium, February 22-25, 2009*.
[39] R. Benadjila, O. Billet, S. Gueron, and M.J. Robshaw. The Intel AES Instructions Set and the SHA-3 Candidates. In *ASIACRYPT '09: Proceedings of the 15th International Conference on the Theory and Application of Cryptology and Information Security*, pages 162–178. Springer-Verlag, 2009.
[40] G. McGraw. *Software Security: Building Security In*. Addison-Wesley, 2006.
[41] W.H.W. Tuttlebee. Software-defined radio: facets of a developing technology. *Personal Communications, IEEE*, 6(2):38–44, Apr. 1999.
[42] W.W. Lu. Open wireless architecture and enhanced performance. *Communications Magazine, IEEE*, 41(6):106 – 107, June 2003.
[43] D.D. Gajski, S. Abdi, A. Gerstlauer, and G. Schirner. *Embedded System Design: Modeling, Synthesis and Verification*. Springer Verlag, 2009.
[44] J. Björkqvist and S. Virtanen. Convergence of hardware and software in platforms for radio technologies. *Communications Magazine, IEEE*, 44(11):52 –57, Nov. 2006.
[45] J. Isoaho, S. Virtanen, and J. Plosila. Current challenges in embedded communication systems. *International Journal of Embedded and Real-Time Communication Systems*, 1(1):1–21, 2010.

 ISBN: 978-1-61208-126-7

Performance of Probabilistic Broadcasting of Dynamic Source Routing Protocol

Muneer Bani Yassein, S. Al-Rushdan,
Computer Science Department
Jordan University of Science and Technology (JUST)
Irbid 22110, Jordan
masadeh@just.edu.jo

Wail Mardini, Yaser Khamayseh
Computer Science Department
Jordan University of Science and Technology (JUST)
Irbid 22110, Jordan
mardini@just.edu.jo, yaser@just.edu.jo

***Abstract* - Blind flooding have been proposed to perform route discovery operations in Mobile Ad-hoc Networks as an early method, but it suffers from a serious problem relied to the broadcast storm problem. Several probabilistic approaches have been proposed to overcome this problem, such as fixed probabilistic, adjusted probabilistic and smart probabilistic schemes. This paper investigates the use of probability with Dynamic Source Routing Protocol (DSR) algorithm to overcome the broadcast storm problem. The paper invistigates issues regarding the implementation and integration of probability in DSR algorithm and how it can be improved. Simulation results show that the new scheme provides good results in performance levels by taking in consideration the status of the network density (sparse versus dense networks).**

Keywords-Source Routing; Probabilistic Flooding; Fixed Probability; Broadcasting

I. INTRODUCTION

In computer networks, the key functionality is to transmit and receive data. Controlling this functionality requires routing the data from the source to the destination. In mobile ad hoc networks, this is a critical and sensitive issue that has to be investigated due to the limitation on such networks, as mobile ad hoc network is designed to be robust. The nodes in such networks are continuously moving and changing there positions, and therefore, the topology of the network are changing frequently, the process of finding the route from source to destination is a challenging task.

One of the first route discovery mechanisms used is blind flooding [1][2]. In blind flooding, when a node receives a Route Request (RREQ) message for the first time, it retransmits it to all its neighbors. This approach is costly and can cause what is called broadcast storm problem [3].

Several schemes have been proposed to overcome this problem using probabilistic approaches. For example, using fixed probability, in which the node rebroadcast the RREQ with a predetermined probability *p*. Other approaches used adjusted or smart probability where the rebroadcasting of the RREQ is performed according to a probability calculated based on some local information of the node's neighbors [4].

In this paper, we investigate the use of probability with DSR algorithm in order to reduce the effects of broadcast storm problem and reduce the contention aiming at improving the performance of the network. The basic idea is to gather useful information about the current status of the network and then make the routing decision based on the collected data.

Finally, for further improvements, we propose a technique to use a global knowledge about the network based on some estimations about the network to enhance the accuracy of the routing decisions.

II. RELATED WORK

One of the earliest techniques used to find a route from a certain source node to a destination node is flooding. Flooding is the basic technique used and implemented in DSR [1] and AODV [2] routing protocols. Flooding is performed by broadcasting a RREQ packet from the source node to all of its neighbors. If one of the neighbors is the destination, it replies by sending a Route Reply (RREP) packet, otherwise, it rebroadcast the request to all it neighbors. If a node (not the destination) receives a RREQ, it will check if the packet has seen before, if so, it will discard the packet, else it will rebroadcast it until the destination node is reached where RREP packet is sent. The concept of flooding is attractive due to its simplicity, effectiveness and ease of implementation.

However, this technique can lead to a serious problem called broadcast storm problem [3]. This problem occurs because the flooding is likely to happen frequently in ad hoc networks. Due to the number of rebroadcast operations, this will result in many redundancies in RREQ packets and cause a contention in the network.

Several approaches were proposed to reduce the redundancy and control rebroadcasting process. One of the earliest techniques is using probabilistic broadcasting in which the intermediate nodes rebroadcast the RREQ according to a certain probability [6].

Bani yassein and Bani Khalaf in [4] used the concept of four probabilities (*P*'s) in their smart probabilistic approach to calculate the rebroadcasting probability based on local information of the neighbors. They choose the value of *p*'s such that *p1>p2>p3>p4* without specifying initial value for *p*.

 ISBN: 978-1-61208-126-7

A. Overview of Dynamic Source Routing Protocol (DSR)

The DSR protocol consists of two phases; route discovery phase which is a mechanism for finding a route from source to destination, and route maintenance phase which is the process of discovering the failure or keeping the link active [1].

When a node creates a packet to be sent to a destination, a route path is appended to the head of packet to determine which path the packet should move through. The node can find this path by searching the route cache for previously learned paths. If the path could not be found in route cache, then this node sends a route discovery packet to find a path to the destination. When a node receives a route discovery packet, it returns a route replay packet to the sender, if it is the target, containing a copy of accumulated route through the network. When the sender receives this message, it put it in its cache to be used in later requests. If the node is not the target node, it appends its address to the route path accumulated in the message and forwards it by a local broadcast packet. If the node finds it own address listed in the message or the message contains a request id that has seen before, it discards this message. The destination node when receiving a route discovery packet, it looks in its route cache to fine a route to the source, if the path found in cache, the destination send pack route replay message through this path, otherwise, the destination starts its own route discovery process.; However, this operation may cause infinite loop, therefore the destination may include the path in the route replay message. Also the destination could reverse the source route if the network has a bidirectional link. The destination saves route replay message in send buffer which contains the messages that cannot be transmitted and try to retransmit the messages later on. Also it can be time stamped so it can be discarded later

B. Adjusted Probabilistic broadcasting

The adjusted probabilistic broadcasting [4] depends on calculating the average number of neighbors in the network. Node's *a* neighbors are defined as the nodes that can be directly reached by node *a*. The average number of neighbors can be calculated as the following:

$$avg = \frac{\sum_{i=1}^{n} N_i}{n} \quad (1)$$

where N_i is the number of neighbors of node *I* and *n* is the number of neighbors in the network. In [4], the nodes estimate the average using the following equation:

$$avg = (N-1)0.8\frac{\pi^2}{A} \quad (2)$$

where *N* is the number of nodes in the network and *A* is the area of the ad hoc network.

Once a node receives a RREQ, it checks if this is not the first time the node receives this RREQ, the node will discard the RREQ, else, it will find the number of its neighbors. If the number of its neighbor is less than *avg*, then the node is probably in a sparse area and therefore high broadcast probability is assigned to the RREQ. On the other hand, if the number of neighbors is greater than *avg*, then the node is probably in dense area and a low probability is assigned to the RREQ.

C. Smart Probabilistic Broadcasting

Smart probabilistic broadcasting algorithm [1] defines four values of probabilities *p1*, *p2*, *p3*, *p4* where *p1>p2>p3>p4*. The algorithm calculates the *avg* as above and calculate the average number of neighbors for the nodes whose number of neighbors is less than *avg* (*avg1*) and the average number of neighbors for the nodes whose number of neighbors is greater then *avg* (*avg2*) using the following equations:

$$avg1 = \frac{\sum_{i=1}^{n} N_i}{k} \quad (3)$$

where *Ni* <avg, *k* is the number of node satisfying the condition, and *n* is the number of nodes in the network.

$$avg2 = \frac{\sum_{i=1}^{n} N_i}{k} \quad (4)$$

where *Ni* ≥ avg, *k* is the number of node satisfying the condition, and *n* is the number of nodes in the network.

Once receiving a RREQ for the first time the node get the values of *avg*, *avg1* and *avg2* then it calculates number of its neighbors *c*. Now, if *c<avg1*, then this node is in low sparse area which means high probability is assigned to RREQ *p=p1*. If *avg1<c< avg*, then the node is in medium sparse area and a medium high probability is assigned to RREQ *p = p2*. If *avg < c < avg2*, then the node is in medium dense area and medium low probability is assigned to RREQ *p=p3*. Finally, if *c>avg2*, then the node is in high dense area and low probability is assigned to RREQ *p=p4*.

III. MOTIVATION

DSR is one of the earliest routing algorithms that have been introduced. It uses flooding technique for route discovery operation. However, as mentioned above, although blind flooding is simple, effective and easy to implement, it can lead to a serious problems like broadcast storm problem.

Many approaches have been proposed to improve the route discovery process. One of the most important approaches is probabilistic broadcasting using fixed probability approaches[4], dynamic probabilistic

 ISBN: 978-1-61208-126-7

approaches, adjusted probabilistic or two-p scheme and smart probabilistic of four-p approaches [3][4].

The aim of using probability is to reduce the redundancy and overlapping in radio signals and hence reduce the contention and collision in the network.

Those techniques mentioned above have been implemented and tested using AODV routing scheme and shows a significant improvement and increase in network performance with the AODV protocol.

As stated in [1], typically, AODV performance is close to the DSR performance. However, in high mobility environments, DSR outperforms AODV because AODV spends significant amount of time to expand the search range. Search in route discovery phase in both AODV and DSR have the same overhead but with a noticeable advantage to DSR as DSR performs less route discovery operations on control overhead over transmission bytes. However, with respect to delivery bytes, AODV outperforms DSR due the increased reliability in AODV.

The probabilistic techniques have not yet investigated using DSR algorithms. A significant improvement is expected to be achieved when implementing probabilistic techniques (fixed, adjusted and smart probabilistic) with DSR.

We propose to use a probabilistic technique in DSR routing scheme based on local knowledge about the neighbors collected by the nodes. The proposed framework will not generate an extra overhead to the network as it will uses the "hello" message mechanism to gather this data. The "hello" message mechanism is explained later in this paper.

IV. IMPLEMENTATION

To implement adjusted and smart probability the node must have some knowledge about the state of the network (its neighbors and their neighbors). Thus, a technique must be defined that enable the node to collect this information in repeated manner. Hence in this work, we define **hello message** to be used with DSR algorithm since the original DSR algorithm was designed to be fully dynamic and did not use hello message. Hello message was defined in our approach to be sent periodically by nodes every some fixed period of time, each hello message contain a broadcast destination address and the originated node source address along with the number of neighbors of the sending node so any node that receives this message will indicate that the sending node is a neighbor and has *n* neighbors and store this information in a local data structure so it can be used when needed, also the hello massage must include a Time To Live (TTL) value equal to 1 to prevent propagating this massage to other nodes other than its neighbors. however not receiving a hello massage for some time this will indicates that a node is no longer a neighbor and has to be discarded from the local data structure, this can be done by adding a time stamp that indicates the time of receiving the hello message and add it to the local data structure along with the sending node and its neighbors, a periodic check is performed based on expiration time property so any neighbor that did not send a hello message for some period of time is no longer a neighbor and has to be eliminated from the local data structure. After collecting information sent by all its neighbors any node in the network can use this information to calculate the probability of resending a route request.

A.Simulation Environment

For the purpose of testing our approach we use NS2 simulator version 2.33 running under Fedora Linux operating system.

B.Simulation Parameters

We use NS2 simulator to test our approach, the simulation area was 250×250 m, 500×500 m, 750×750 m and 1000×1000m the node bandwidth is 2Mbps using IEEE802.11 MAC layer protocol. Regarding the number of nodes the approach was simulated with 25, 50, 75 and 100 mobile node. Nodes speed was chosen to be 4, 8, 12 and 16 m/s [5]. And the last parameter is the number of connections used the simulation is performed at 5, 10, 15 and 20 connections.

In order to test our approach several scenarios must be generated to represent different simulation parameters, each point at the simulation was tested using 10 deferent scenarios to ensure an acceptable degree of confidence, also different algorithms was tested using the same scenarios at the same point in the simulation, and using different scenarios at different points in the simulation to represent different simulation parameters.

C. Evaluation Criteria

The following metrics where used as an evaluation criteria's:

1- Average end-to-end delay: it include all end to end delay in data packet sending from source to destination which include delay caused by route discovery, MAC layer delay and application layer delay.
2- Routing overhead: it is the number of route request packets generated through the simulation process to find a path for sending a data packet.
3- Normalized routing load: it is the ratio of routing packet transmitted to the data packet delivered through the simulation.

D. Average End to end Delay

Figure 1 shows the average end-to-end delay for Blind Flooding DSR, Fixed Probability DSR, 2P DSR and 4P DSR against the simulation area. The results were obtained

 ISBN: 978-1-61208-126-7

for simulation area between 250×250 and 1000×1000 with 50 node, speed of 16m/s and number of connections 10. It can be seen that the for areas below 750x750 the delay is under 0.2 whereas the delay is increased rapidly afterwards at 1000x1000 to around 1.6.

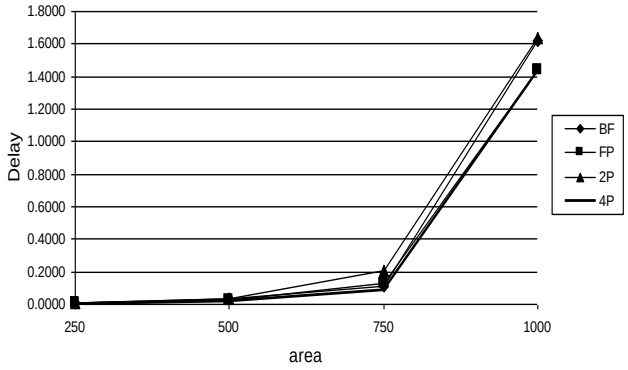


Figure 1. End-to-end delay against area when n=50, speed 16m/s and connections=10.

At simulation area of 250×250, the results from the four algorithms seems to be identical but as the simulation area increase the 4P and FP show better performance than BF and 2P.

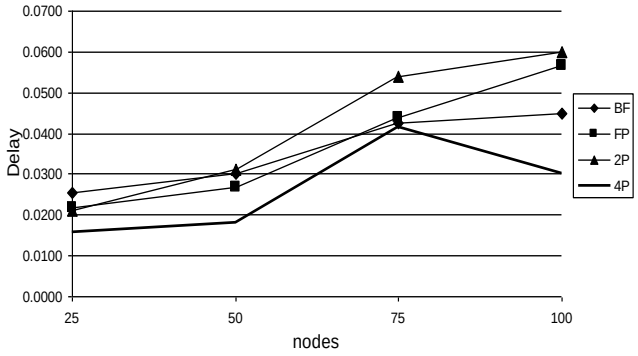


Figure 2. End-to-end delay against number of nodes when area= 500×500, speed 16m/s and connections=10.

Figure 2 shows the end-to-end delay against the number of nodes for the four algorithms at simulation area 500×500m, a speed of 16m/s and number of connections equal 10.

As can be seen that the 4P algorithm gives better performance than the other algorithms because of its flexibility in determining different probabilities, while the BF still shows a better improvement than 2P and FP.

Figure 3 shows the end-to-end delay against the speed of nodes for the four algorithms at simulation area 500×500m, a number of nodes of 50 and number of connections equal 10.

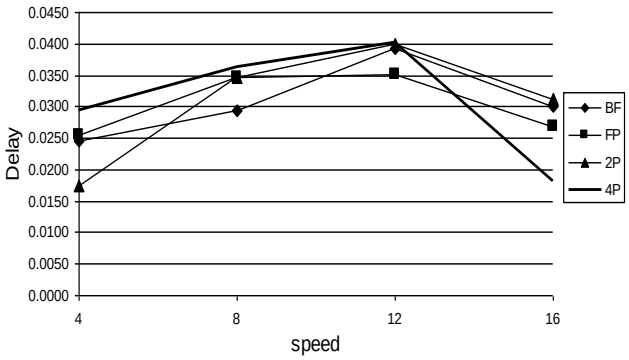


Figure 3. End-to-end delay against the speed of nodes when area=500×500, n=50 and connections=10.

As the speed of nods increases, the 4P algorithm shows a significant improvement over the other algorithms; moreover, FP algorithm performance is better than BF and 2P.

V. "HELLO" PACKETS

"Hello" packets are periodically generated by a given node in order to know the number of its neighbors. These packets are extra control packets sent by nodes to successfully accomplish broadcast operations. Each node sends a short packet that informs its neighbors of its presence. So, a node can know its neighbors by simply listening to the medium. Since nodes obtain neighborhood information through "Hello" packets, the information in the "Hello" packet varies depending on it usage. Thus it is necessary to quantitatively assess the impact of the size of the "Hello" packets on the overhead involved and thus be able to comment on any possible performance tradeoffs. To this end, we have used a "Hello" packet with a size of 12 bytes for exchanging neighborhood information.

VI. CONCLUSION AND FUTURE WORK

The paper presented a performance evaluation of using probabilistic approach in DSR routing algorithms. It is noticeable that the performance of probability approach is better than Blind Flooding especially at high number of connections and high simulation area. Also it is noticeable that at low network area (e.g., 250×250 and 500×500), all the four algorithms show relatively close performance. But at higher network area (e.g., 750×750 and 1000×1000) probabilistic approaches shows a significant improvement over the Blind flooding approach.

The simulation experiments show that our proposed scheme significantly outperforms the 2P-based and flooding schemes in terms of reducing overhead at low speed of nodes (1 m/s). In addition, the 4P algorithm substantially outperforms the other algorithms in terms of reducing average end-to-end delay for low speed nodes. Regarding

 ISBN: 978-1-61208-126-7

packet delivery ratio, the experimental results show that our scheme outperforms the other schemes.

REFERENCES

[1] David B. Johnson, David A. Maltz, and Josh Broch. "DSR: The dynamic source routing protocol for Multi-Hop wireless ad hoc networks". In Ad Hoc Networking, edited by Charles E. Perkins, Chapter 5, pp. 139–172, 2001. http://www.monarch.cs.rice.edu/monarch-papers/dsr-chapter00.pdf (last accessed: Auguest 15, 2010).

[2] C. E. Perkins and E. M. Royer. Ad-hoc on-demand distance vector routing. In Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99. Second IEEE Workshop on, pp. 90–100, 1999.

[3] Sze-Yao Ni, Yu-Chee Tseng, Yuh-Shyan Chen, and Jang-Ping Sheu. The broadcast storm problem in a mobile ad hoc network. In MobiCom '99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking, pp. 151–162, New York, NY, USA, 1999. ACM.

[4] Muneer Bani Yassein, Mustafa Bani Khalaf, and Ahmed Al-Dubai. A new probabilistic broadcasting scheme for mobile ad hoc on-demand distance vector (AODV) routed networks. The Journal of Supercomputing, March 2010, pp. 196-211.

[5] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. Wireless Communications & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, 2(5): pp. 483–502, 2002.

[6] E. M. Royer and Chai-Keong Toh. A review of current routing protocols for ad hoc mobile wireless networks. Personal Communications, IEEE, 6(2): pp. 46–55, August 2002.

 ISBN: 978-1-61208-126-7

Sensitivity Analysis of Availability of Redundancy in Computer Networks

Rubens de S. Matos Júnior*, Almir P. Guimarães*†,
Kadna M. A. Camboim*, Paulo R. M. Maciel*
*Center of Informatics
Federal University of Pernambuco
Recife, Brazil
Email: {rsmj,apg2,kmac,prmm}@cin.ufpe.br
†Campus Arapiraca
Federal University of Alagoas
Arapiraca, Brazil
Email: almir@arapiraca.ufal.br

Kishor S. Trivedi‡
‡Dept. of Electrical & Computer Eng.
Duke University
Durham, USA
Email: kst@ee.duke.edu

***Abstract*—In this paper, we investigate the availability modeling of computer networks with redundancy mechanisms. Sensitivity analysis is applied in order to find the bottlenecks of system availability. We use Markov chains for the analytical evaluation of complex scenarios. We apply our proposed modeling approach in a case study to evaluate how sensitive dependability is to failure and recovery times of different components in an enterprise network. The influence of network topologies is also considered in our case study.**

***Keywords*—Availability; Markov Chains; Sensitivity analysis; Computer networks**

I. Introduction

Over the last few years, the use of data networks has significantly increased. This considerable growth is somewhat related to the convergence of many different services on the same transmission technology. These services should be continuously provided even when events like congestion, link failures, routing instabilities, sabotage, natural disasters, hardware or software failures happen. The design, deployment and management of communication network infrastructure ought to meet such requirements. The possibility of identifying points where the unavailability or downtime of these networks may put the business at risk is an interesting track to be followed by organizations.

Recently, much has been done to deal with issues relating to the availability of computer networks. Researchers have used different approaches to deal with these problems, including sensitivity analysis techniques and the development of advanced redundancy mechanisms.

Zou *et al.* [1] discusses algorithmic methods to compute network availability for a given topology and presents two tools for computation of network availability in large and complex networks. Semaan [2] discusses different issues related to network availability. First, the paper presents some of the elements that impact the availability of a solution. Then, it discusses how network designers can calculate the exact availability of their solution and provides means to determine the optimal level of availability. Trivedi *et al.* [3] presents a new classification of dependability and security models for systems and networks. It also presents several individual model types such as availability, confidentiality, integrity, performance, reliability, survivability, safety and maintainability models. Furthermore, it is shown that individual model types can be combined to form composite dependability model types. The dependability/security models can be represented as combinatorial models, state-space models, and hierarchical models.

In this paper, we focus on the availability of data networks, including redundancy mechanisms. Several scenarios are evaluated through analytic-numeric solution of Markov chains [4]. The model parameters used were obtained from manufacturers of network elements, as also from experimental measurements. We evaluate the impact of different component parameters on the overall system availability, by means of differential sensitivity analysis.

The rest of the paper is organized as follows: Section II presents basics of availability of computer networks and sensitivity analysis. Section III describes the proposed availability models. Section IV presents the evaluation of availability and its sensitivity for all the proposed models. Finally, Section V discusses the results of this study and introduces ideas for future research.

II. Fundamental concepts

A. Dependability Requirements for Voice and Data Networks

Standard IP applications traffic is characterized by burstiness. However, such applications are not highly sensitive to delay and jitter. On the other hand, voice applications run continuously and steady, they could thus be strongly affected by long delays and jitter [5]. Providing high quality voice service on IP networks is one of the most pressing issues faced by the VoIP community [6].

Critical services, such as VoIP, have strict QoS requirements for both performance tolerance and service dependability. Dependability of a computer system must be understood as the ability to avoid service failures that are more frequent and more severe than is acceptable [7]. Dependability attributes

 ISBN: 978-1-61208-126-7

include the concepts of availability, reliability, safety, integrity and maintainability [7].

Inputs to availability models include component Mean Times to Failure (MTTF) and Mean Times To Repair (MTTR). The hardware component MTTFs are generally supplied by the manufacturer. The MTTRs are tightly related to the maintenance policy adopted by the organization.

B. Parametric Sensitivity Analysis

Sensitivity analysis is a method of determining the most influential factors on model results [8], [9]. The effect of changes in data distribution function and the impact of changes in parameter values are examples of study subjects for sensitivity analysis. When dealing with analytic models such as Markov chains, parametric sensitivity analysis is a particularly important technique for assessing the effect of changes in the rate constants on the measures of interest. This approach may be used to find performance or availability bottlenecks in the system, thus guiding an improvement and optimization [10].

There are many ways of conducting sensitivity analyses. The simplest method is to repeatedly vary one parameter at a time, while keeping the others fixed. When applying this method, the sensitivity ranking is obtained by noting the corresponding changes in the model output. Other techniques include factorial experimental design [11], correlation analysis, regression analysis and perturbation analysis (PA). Differential analysis, also referred to as parametric sensitivity analysis or the direct method, is the backbone of nearly all other sensitivity analysis techniques [9]. This method is chosen in this paper, as it can be performed in an efficient computational manner on analytic models commonly used in performance and availability analyses.

Parametric sensitivity analysis is performed by computing the partial derivatives of the measure of interest with respect to each input parameter. For instance, the sensitivity of a given measure Y, which depends on a parameter λ, is computed as in Equation (1), or (2) for a scaled sensitivity.

$$S_\lambda(Y) = \frac{\partial Y}{\partial \lambda} \tag{1}$$

$$S^*_\lambda(Y) = \frac{\partial Y}{\partial \lambda}\left(\frac{\lambda}{Y}\right) \tag{2}$$

A number of researchers have already demonstrated how to perform parametric sensitivity analysis in a variety of analytic models. In [10], the basics of transient sensitivity analysis in continuous time Markov chains (CTMC) are presented. Sensitivity functions for Markov chains were recently implemented in the SHARPE package [12], making use of the techniques described in the papers we have just cited. Since the reduced reachability graph of a Stochastic Petri Net (SPN) is a Markov chain, this kind of model may also be analyzed, by following the steps indicated in [13]. Their work includes the implementation of sensitivity analysis features in the SPNP package [14]. Queueing systems are another example of analytic models whose sensitivity analysis has been described in [15].

III. Proposed Availability Models

Traditional evaluation techniques for availability use Markov chains and Markov reward models. In this section, we present three CTMC (Continuous Time Markov Chain) availability models (Figures 4, 5 and 6). The first one represents a system without any redundancy. The second one represents a system with aspects of fault-tolerance based on link redundancy (see Figure 2). Then, the last one represents a system with aspects of fault-tolerance based on warm-standby redundancy (see Figure 3). This approach is characterized by fault detection and recovery mechanisms. The dependability models can be evaluated using tools such as SHARPE (Symbolic Hierarchical Automated Reliability and Performance Evaluator) [12].

A. Platform Description

The following three scenarios were used as a basis for the availability analytic models presented in this paper. They also served as experimental testbeds, from which some failure and recovery parameters were obtained, as well as to validate the analytic results obtained from the respective Markov chains.

1) First and Second Scenarios: In the first scenario, the testbed is composed of two machines, a switch and two routers that are connected by a single link (see Figure 1). In the second scenario, the testbed is composed of two machines, a switch and two routers that are connected by redundant links (L0 and L1 - see Figure 2). When the main link (L0) fails, the spare link (L1) assumes the role of the main one. After main link restoration, the system returns to the initial condition.

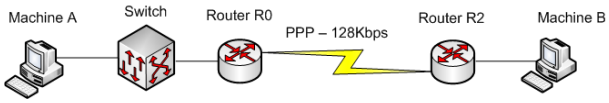


Figure 1. Test Bed - Scenario 1.

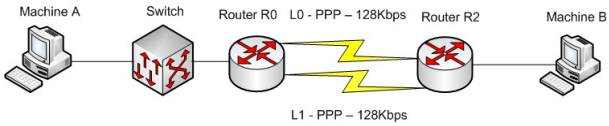


Figure 2. Test Bed - Scenario 2.

2) Third Scenario: In this scenario, the testbed is composed of two machines, a switch and three routers (see Figure 3). The system uses fault-tolerance based on warm-standby redundancy. When one of the primary components (R0 or L0) fails, the spare components (R1 and L1) assume the role of the primary components. This switchover process takes time for the spare components to start operation, named Mean Time to Activate (MTTA). After restoration of the primary components, the system returns to the initial condition.

B. CTMC Availability Model without redundancy

In Figure 4, the Markov chain represents the first scenario, which is the simplest one, with no redundancy. There is only

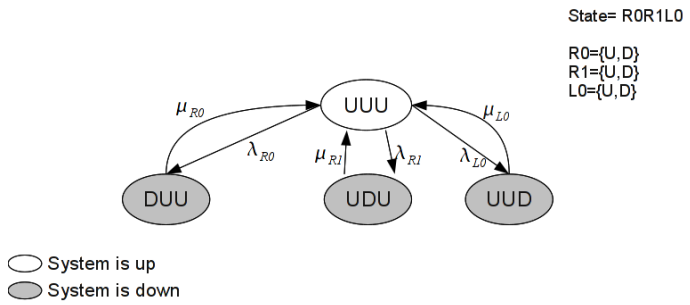


Figure 4. Markov chain for the availability of non-redundant network

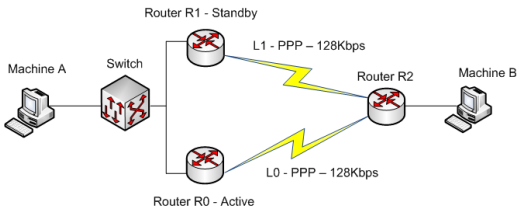


Figure 3. Test Bed - Scenario 3.

TABLE I
STATES OF CTMC MODEL WITH LINK REDUNDANCY

State	Description
UUU	The System is UP
DUU	Down state, Router R0 failed
UDU	Down state, Router R1 failed
UUD	Down state, Link L0 failed

one link, named L0, connecting router R0 and router R1. In this model, the normal operation of a component is denoted by the label U (up), and a failed component is represented by label D (down). A state in the Markov chain is defined by a sequence of labels, representing router R0, router R1 and link L0, respectively. We assume the failure and repair time of each component are exponentially distributed. λ_{R0}, λ_{R1} and λ_{L0} are the respective failure rates of R0, R1 and L0. In a similar notation, μ_{L0}, μ_{R0} and μ_{R1} are the respective repair rates of each system component. Once any component (R0, R1, or L0) has failed, the overall system is in a down state and subsequently no additional failures occur until the component is repaired, so that the expansion of state space stops at the first down state. In Table I, a description of each state is given. For this model, the system is up and running only in the state UUU. All the other states are shaded gray in Figure 4, representing the system down states.

C. CTMC Availability Model with link redundancy

In Figure 5, we consider a system that has redundancy only at the link level, as illustrated in Figure 2. The normal operation of a component is denoted by the label U (up), and a failed component is represented by label D (down). A state in the Markov chain is also defined by a sequence of labels, representing router R0, router R1, link L0, and link L1, respectively. The ideal condition for this system is denoted by state UUUU, in which all components are in non-failed condition. Failure transitions have rates λ_X, and repair transitions have rates μ_X, where $X \in \{R0, R1, L0, L1\}$, representing each system component. In states shaded gray, the system has failed, due to a failure in one of the routers, or a failure in both links. We assume that in those states no additional failures occur in the remaining components, since they are in an idle condition. Another assumption for this model is that there is a repair policy, that prioritizes the repair of link L0 over link L1 when both are failed. We do not consider any priority in the repair of routers because it is not possible in this model to have both routers down, since a failure in any of them brings the overall system to a down state. In Table II, a description for each state is given.

TABLE II
STATES OF CTMC MODEL WITH LINK REDUNDANCY

State	Description
UUUU	The System is UP
DUUU	Down state, Router R0 failed
UDUU	Down state, Router R1 failed
UUDU	The System is UP, Link L0 failed
UDDU	Down state, R1 and L0 failed
DUDU	Down state, R0 and L0 failed
UUDD	Down state, L0 and L1 failed
UUUD	The System is UP, L1 failed
DUUD	Down state, R0 and L1 failed
UDUD	Down state, R1 and L1 failed

D. CTMC Availability Model with router redundancy

We present in Figure 6 a Markov chain that represents the system illustrated in Figure 3. The failure and repair rates of each component are represented by λ_X and μ_X, where $X \in \{R0, R1, R2, L0, L1\}$. Rates α_R and α_L are the inverse of mean time to activate the spare router and the spare link, respectively.

For simplicity, we have made some simplifications that do not significantly affect the results we obtain from the analysis. One of the assumptions for this model is that there is a priority in the repair of components. Router R2 has the higher priority, followed by router R0, link L0, router R1, and link L1, in descending order. We also consider that no failure is possible

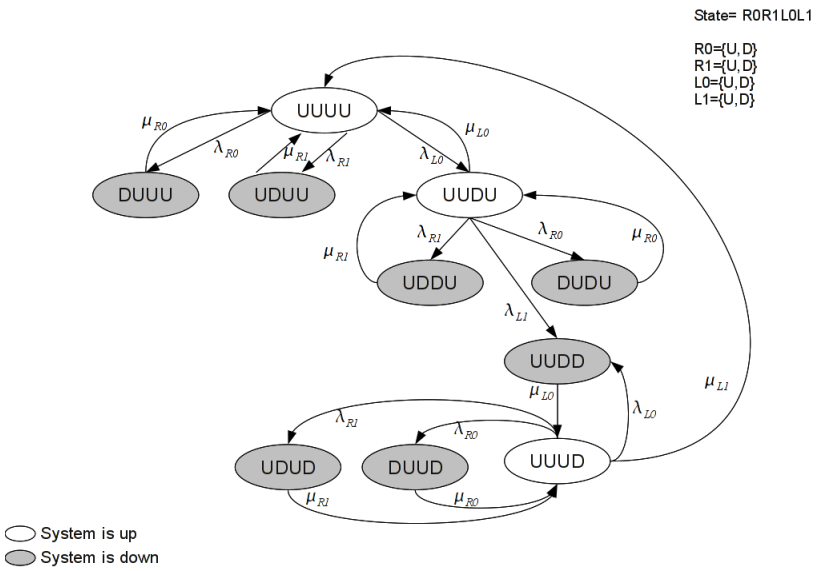


Figure 5. Markov chain for the availability of link-redundant network

when a component is in waiting condition.

The nomenclature used for the states in this model is based on the current condition of each system component, in the following order: router R0, link L0, router R2, router R1, link L1. A letter U indicates the up condition, when component is active. Letter D denotes a down condition for that component, meaning that it has failed, and that a repair is needed. Letter W represents a waiting condition, in which the component is not being used, but is ready to enter in active mode, as soon as it is needed. Therefore, state UUUWW denotes router R0, link L0, and router R2 are active, router R1 and link L1 are on waiting condition. For an active state, the system must present one of the following combinations: UUUWW, DWUUU, WDUUU, UUUDW or UUUWD (see Figure 6). Particularly, in the DWUUU state, router R2 with spare router and link (R1 and L1) are active, while R0 is in down state and L0 is waiting, since it only works together with R0. A similar situation happens in WDUUU state, where R2, R1 and L1 are active, but L0 is down, leaving R0 in a waiting condition.

Figure 6 shows the initial state of the system, UUUWW. With rate λ_{R0} a failure happens and brings the system to down state (DUUWW). Similarly, with rate λ_{L0} a failure happens and brings the system to down state (UDUWW). Likewise, with rate λ_{R2} a failure happens and brings the system to down state (UUDWW). In this case, the repair with rate μ_{R2} brings the system back to state (UUUWW).

After detecting a failure, a switchover occur making the spare components (R1 and L1) active. In the states DUUWW and UDUWW, the system activates the spare components with rates α_R and α_L corresponding to router R0 and link L0 failures, respectively. After switch-over, the standby components are able to take over the failed components, bringing the system to and active state (DWUUU or WDUUU). The repair with rates μ_{R0} and μ_{L0} bring the system back to initial state. Before a repair happens, another failure with rate λ_{R1}, λ_{L1} or λ_{R2} may bring the system to a down state (DWUDU, DWUUD, DWDUU, WDDUU, WDUDU, WDUUD). From states DWDUU and WDDUU, with rate μ_{R2}, the system comes back to active states (DWUUU and WDUUU). Similarly, from states DWUDU and DWUUD, with rate μ_{R0}, the system comes back to active states (UUUDW and UUUWD). Likewise, from states WDUDU and WDUUD, with rate μ_{L0}, the system return to active states (UUUDW and UUUWD). From the active states, the repair with rate μ_{R1} or μ_{L1} brings the system to initial state. Finally, from UUUDW and UUUWD, the system can return to down state with rate λ_{R0}, λ_{L0} or λ_{R2}. The repair with rate μ_{R0}, μ_{L0} or μ_{R2} brings the system back to active states (UUUDW and UUUWD).

In Table III, we see the system availability condition for each state of this Markov chain. Note that only on 5 states the system is operational: UUUWW, DWUUU, WDUUU, UUUDW, UUUWD.

IV. Case Study

We concentrate our attention on parametric sensitivity analysis, as a technique to compute the effect of changes in the rate constants of a Markov model on the measures of interest. Parametric sensitivity analysis helps: (1) to guide system optimization, (2) to find availability, performance, and performability bottlenecks in the system, and (3) to identify the model parameters that may produce significant modeling errors. In this paper, (1) and (2) are the main purposes, although the identification of errors in the early versions of the proposed models was also possible through such analysis.

We consider the testbed shown in Section III-A to perform a parametric sensitivity analysis using the proposed dependability CTMC models. The MTTFs of components used in this work are respectively: 131,000 hours for routers and 11,988

 ISBN: 978-1-61208-126-7

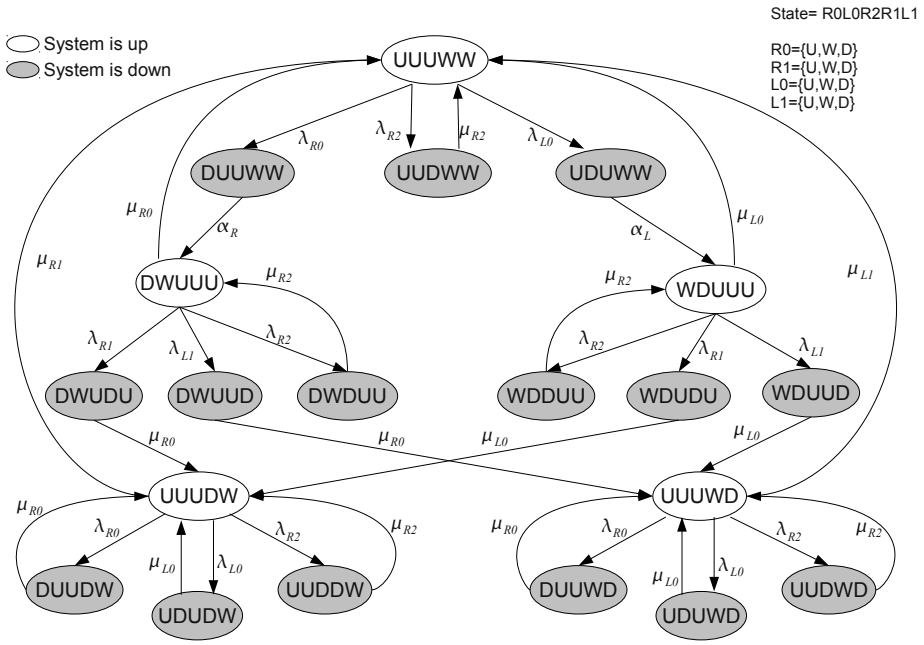


Figure 6. Markov chain for the availability of router-redundant network

TABLE III
STATES OF CTMC MODEL WITH ROUTER REDUNDANCY

State	Description
UUUWW	The System is UP
DUUWW	Down state, Switchover Started
UDUWW	Down state, Switchover Started
UUDWW	The System is Down
DWUUU	The System is UP
WDUUU	The System is UP
DWUDU	The System is Down
DWUUD	The System is Down
DWDUU	The System is Down
WDDUU	The System is Down
WDUDU	The System is Down
WDUUD	The System is Down
UUUDW	The System is UP
UUUWD	The System is UP
DUUDW	The System is Down
UDUDW	The System is Down
UUDDW	The System is Down
DUUWD	The System is Down
UDUWD	The System is Down
UUDWD	The System is Down

hours for links. We use a mean time to repair (MTTR) equal to 12 hours, for all components. Those values shall be considered as the base case throughout this section, unless another value is specified in each specific analysis. Notice that all λ_i in the models of previous section are equal to $1/MTTF_i$, and all μ_i are equal to $1/MTTR_i$. Sensitivity analysis of steady-state availability is carried out by computing $S^*_{MTTF_i}(A)$ as the scaled sensitivity of availability with respect to $MTTF_i$, and $S^*_{MTTR_i}(A)$ as the corresponding measure with respect to $MTTR_i$.

In the base case, using the values we have just mentioned, the steady-state availability for the first scenario is 0.998817194. The second scenario, in which link redundancy is added, presents an availability of 0.999815827. In the third scenario, which has redundant router, the availability increases to 0.999906968.

Initially, we consider the sensitivity analysis regarding the third scenario, in Section III-A2. The values for $S^*_k(A)$, where A is the system steady-state availability and k is each of the components' MTTF and MTTR, were computed using sensitivity analysis features developed for the SHARPE package. In Table IV, we see that parameters $MTTF_{R2}$ and $MTTR_{R2}$ assume the greatest importance in system steady-state availability, since they have the highest sensitivity values. Any change in these parameters will have a major impact on system availability, but in opposite directions. Sensitivity with respect to $MTTF_{R2}$ is positive, since the availability increases when this parameter increases. In contrast, $S^*_{MTTR_{R2}}(A)$ is negative, because a smaller repair time of R2 implies an increased availability. In Table IV, we can also notice that time to repair spare components ($MTTR_{R1}$ and $MTTR_{L1}$) are the ones with smallest impact on the system availability. This result matches the results from the established repair policy, since failed spare components are repaired only after main components have returned to normal operation.

Figure 7 depicts a plot for the system availability, in which the MTTF parameters for the system links (L0 and L1) were changed one at a time, and the analytic model of Figure 6 was solved. This plot confirms that efforts in expanding the time to failure of link L0 have more impact on system availability than increases in $MTTF_{L1}$ do. Figure 8 also validates the

TABLE IV
SENSITIVITY OF AVAILABILITY FOR SCENARIO WITH ROUTER REDUNDANCY

Parameter k	$S_k^*(A)$
$MTTF_{R2}$	9.15946628e-05
$MTTR_{R2}$	-9.15946627e-05
$MTTR_{L0}$	-2.18237573e-06
$MTTF_{L0}$	1.31711540e-06
$MTTF_{L1}$	1.09121018e-06
$MTTR_{R0}$	-1.99712368e-07
$MTTF_{R0}$	1.20531140e-07
$MTTF_{R1}$	9.98582261e-08
$MTTR_{L1}$	-1.19080014e-09
$MTTR_{R1}$	-1.08971849e-10

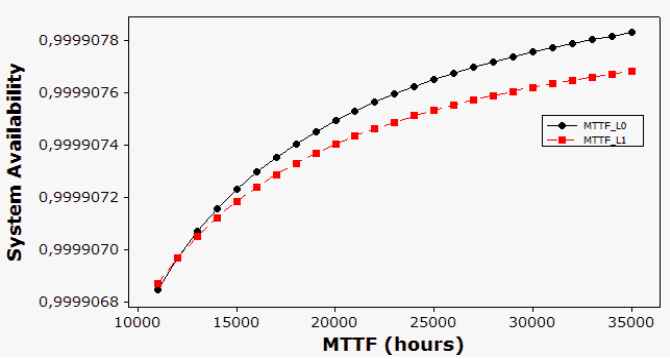


Figure 7. Effect of each link MTTF on system availability (scenario 3)

results from sensitivity ranking. If we make $MTTF_{R2}$ bigger the benefits will be much higher than that resulting from enhancements on either R0 or R1 MTTFs.

MTTR is an important factor for system availability since it will affect the downtime of network elements. For a redundant topology (third scenario), Figure 9 shows the system availability as a function of each component MTTR. Router R2 is the component whose time to repair causes the biggest effect on the steady-state availability, followed by link L0. This information can also be obtained comparing the corresponding values for each MTTR in Table IV.

Then, we will analyze the impact of R0 and L0 MTTFs on the system availability in each proposed scenarios. The idea is to observe the real impact of these parameters with respect

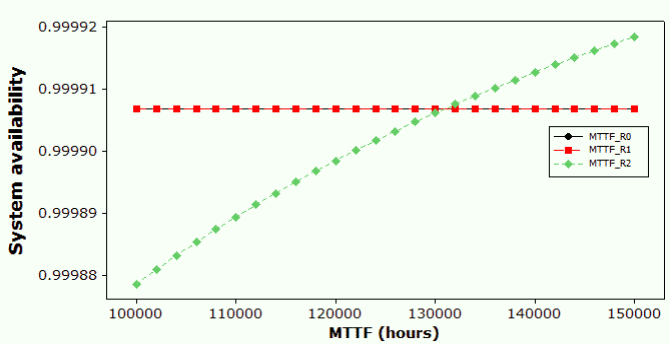


Figure 8. Effect of each router MTTF on system availability (scenario 3)

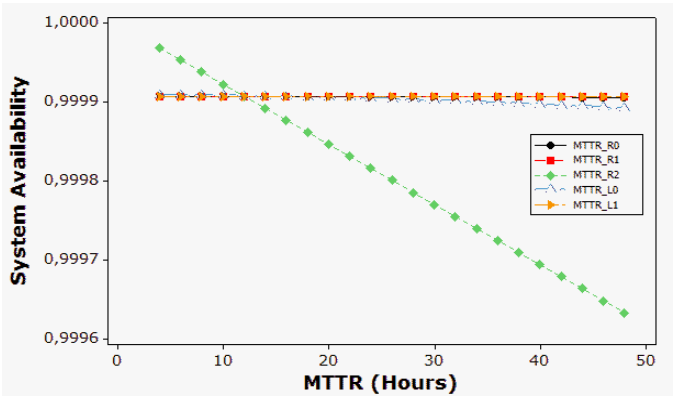


Figure 9. Effect of each component MTTR on system availability (scenario 3)

to system availability in different redundancy schemes (i.e. different scenarios).

Comparing Tables V and VI - also obtained through SHARPE's sensitivity analysis features - we see that the redundancy mechanism makes the system availability less sensitive to failures of primary link (L0), while the sensitivity with respect to $MTTF_{R0}$ is almost the same. The results from differential sensitivity analysis can also be confirmed by comparing, in a scatter plot, the effect of changes in $MTTF_{L0}$ and $MTTF_{R0}$ for each one of our three scenarios. In Figures 10 and 11, we see that in second and third scenarios availability is less affected by increases in $MTTF_{L0}$ and in $MTTF_{R0}$. As in the previous plots, those values were obtained by repeatedly varying the value of each parameter at a time, and solving the Markov chain for that configuration. It is important to state that this comparison using scatter plots becomes more difficult as the number of parameters in the model increases. The ranking obtained from differential sensitivity analysis allows a direct view of the importance order of all parameters.

The reduction of main router impact on availability is confirmed in Table IV, that shows a higher sensitivity with respect to the failure of spare link ($MTTF_{L1}$) than with respect to main router failure ($MTTF_{R0}$). So, actions to increase the MTTF of link L1 should be considered more important than additional efforts to enhance the MTTF of router R0. This kind of decision could not be easily made without an accurate sensitivity analysis, as we have performed in this case study.

V. CONCLUSION

In this paper, we proposed Markov chain models to evaluate several dependability aspects of computer networks in different scenarios. The models support the analysis of system availability along with its services, based on different topologies, redundancy mechanisms and network elements. Sensitivity analysis was applied in order to guide system optimization in terms of steady-state availability.

For future work, we plan to extend these models to include network availability with redundant topologies, different re-

 ISBN: 978-1-61208-126-7

TABLE V
SENSITIVITY OF AVAILABILITY FOR SCENARIO WITH LINK REDUNDANCY

Parameter k	$S_k^*(A)$
$MTTF_{R0}$	9.15861826e-05
$MTTF_{R1}$	9.15861826e-05
$MTTR_{R0}$	-9.15861826e-05
$MTTR_{R1}$	-9.15861826e-05
$MTTF_{L0}$	1.00081664e-06
$MTTR_{L0}$	-1.99963265e-06
$MTTF_{L1}$	9.99815827e-07
$MTTR_{L1}$	-9.99815827e-10

TABLE VI
SENSITIVITY OF AVAILABILITY FOR SCENARIO WITH NO REDUNDANCY

Parameter k	$S_k^*(A)$
$MTTF_{L0}$	9.99817011e-04
$MTTR_{L0}$	-9.99817011e-04
$MTTF_{R0}$	9.14947048e-05
$MTTR_{R0}$	-9.14947048e-05
$MTTF_{R1}$	9.14947048e-05
$MTTR_{R1}$	-9.14947048e-05

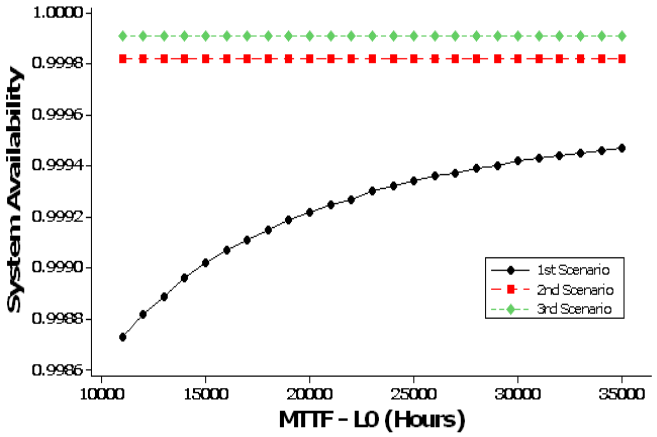


Figure 10. Availability Vs. $MTTF_{L0}$

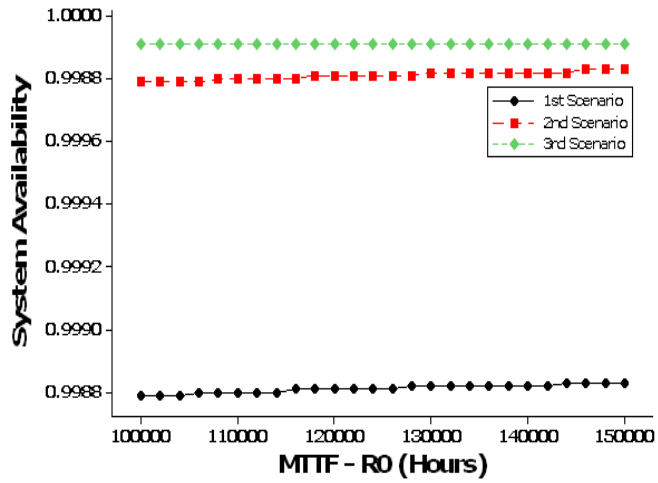


Figure 11. Availability Vs. $MTTF_{R0}$

covery strategies as well as taking into account different failure modes.

REFERENCES

[1] W. Zou, M. Janic, R. Kooij, and F. Kuipers, *On the availability of networks*, in Proc. of BroadBand Europe 2007, Antwerp, Dec. 2007.

[2] G. Semaan, *Designing networks with the optimal availability*, in National Fiber Optic Engineers Conference, Optical Society of America, pp. 1–6, 2008.

[3] K. Trivedi, D. Kim, A. Roy, and D. Medhi, *Dependability and security models*, Proc. DRCN, pp. 11–20, 2009.

[4] Kolmogorov, A. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung (in German). Mathematische Annalen. Springer-Verlag, 1931.

[5] C. Hoene, B. Rathke, and A. Wolisz, *On the importance of a VoIP packet*, In Proc. Of ISCA Tutorial and Research Workshop on the Auditory Quality of Systems, 2003.

[6] H. Sze, S. Liew, J. Lee and D. Yip, *A multiplexing scheme for h.323 voice-over-ip applications*, IEEE J. Select. Areas Commun., vol. 20, no. 7, pp. 1360–1368, 2002.

[7] A. Avizienis, J. Laprie, B. Randell, and C. Landwehr, *Basic concepts and taxonomy of dependable and secure computing*, IEEE Transactions on Dependable and Secure Computing, vol. 1, no. 1, pp. 11–33, 2004.

[8] P. Frank, *Introduction to System Sensitivity Theory*. Academic Press Inc, Sept. 1978.

[9] D. Hamby, *A review of techniques for parameter sensitivity analysis of environmental models*, Environmental Monitoring and Assessment, pp. 135–154, 1994.

[10] J. Blake, A. Reibman, and K. Trivedi, *Sensitivity analysis of reliability and performability measures for multiprocessor systems*, in SIGMETRICS '88: Proceedings of the 1988 ACM SIGMETRICS conference on Measurement and modeling of computer systems.New York, NY, USA: ACM, 1988, pp. 177–186.

[11] R. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling*, John Wiley, 1991.

[12] K. Trivedi and R. Sahner, *Sharpe at the age of twenty two*, SIGMETRICS Perform. Eval. Rev., vol. 36, no. 4, pp. 52–57, 2009.

[13] J. Muppala and K. Trivedi, *GSPN models: sensitivity analysis and applications*, in ACM-SE 28: Proceedings of the 28th annual Southeast regional conference.New York, NY, USA: ACM, 1990, pp. 25–33.

[14] C. Hirel, B. Tu, and K. Trivedi, *SPNP: Stochastic Petri Nets. version 6.0*, Mar. 2010. [Online]. Available: http://www.ee.duke.edu/~chirel/PAPER/paperSpnp.pdf

[15] B. Yin, G. Dai, Y. Li, and H. Xi, *Sensitivity analysis and estimates of the performance for M/G/1 queueing systems*, Performance Evaluation, vol. 64, no. 4, pp. 347–356, 2007.

 ISBN: 978-1-61208-126-7

The Method of Mission Reliability Allocation for Complex System Based on Simulation

Zhao Guangyan, Qin Tong, Sun Yufeng, Hu Weiwei
School of Reliability and System Engineering
Beihang University
Beijing, China
zhaoguangyan@buaa.edu.cn, qintong@dse.buaa.edu.cn, syf@buaa.edu.cn, hww@buaa.edu.cn

***Abstract*—Reliability has become a greater concern in recent years, because high-tech industrial processes with ever increasing levels of sophistication comprise most engineering systems today. Reliability Allocation Method currently used has corresponding advantages in different aspects, but there are still some limitations in practical engineering applications. It is difficult to establish the mission reliability models for large and complex system while facing the degradation use in function, dynamic correlation between structural composition and faults not meeting the independent assumptions, and other cases. Based on the analysis of the advantages and disadvantages of the traditional reliability allocation methods, this paper described in details the mission reliability allocation method of complex system by applying simulation. It can not only conduct a great deal of iteration calculation but also can weigh different allocation schemes without adding additional work, using simulation calculations after the initial establishment of system simulation models. This is suitable for the mission reliability allocation of complex systems in engineering projects. Finally, this paper provided practical cases applying this method into achieving mission reliability allocation of a warship system.**

Keywords- Reliability Allocation; Complex System; Simulation; Model.

I. INTRODUCTION

Reliability has become an ever greater concern in recent years, because high-tech industrial processes with ever increasing levels of sophistication comprise most engineering systems today. Reliability allocation is an important link of system reliability design. Reliability allocation is to allocate reliability indicators provided by the system to subsystems, parts and components according to certain principles and procedures so as to enable designers at all levels to identify its reliability design requirements, accurately estimate manpower, time and resources according to the requirements and study the possibility of achieving this requirement and methods. It is a decomposition process from the integral to part, from big to small and from top to bottom. Reliability allocation can be used in the design of both hardware system and software. While most of literatures present how to do the reliability allocation for hardware design a method of reliability allocation for software and network is described in [1, 2].

Reliability allocation includes basic reliability allocation and mission reliability allocation. Currently, reliability allocation is carried out mainly by constrained system reliability allocation method and unconstrained reliability allocation method. The former mainly includes Lagrange multiplier method, dynamic programming method, and direct search method; while the latter mainly includes equal allocation method, combination in proportion technique and score allocation method [3], and it also includes some methods deriving from above for the score is more rational, such as Fuzzy Decision Method [4,5], AHP [6], using the maximal entropy ordered weighted averaging method [7], Analytical Target Cascading Method [8], sensitivity evaluation method [9] and so on. Unconstrained allocation method is frequently used in engineering. The use of all these methods often needs to establish the reliability model. It is very easy to build the basic reliability model of the system, such as series model; However, it is often difficult to establish the mission reliability model for large and complex system due to complexity of its structure, complexity in functions to be achieved, degradation use in function and dynamic correlation between structural composition, faults not meeting the independent assumptions, and other cases. This results in difficulties in the use of above methods. In this case, this paper, based on simulation ideas [10,11], puts forward the simulation-based reliability allocation method in order to achieve the mission reliability allocation for complex system. This method establishes not the mission reliability model but the simulation logic model of the system by defectively using functional logical relationship between product units, which can be widely applied in engineering.

In this paper, Section Ⅱ introduces the overview of reliability allocation method frequently used. And then the simulation-based mission reliability allocation for complex system, including allocation idea and allocation process are detailed in Section Ⅲ. At last, it is a case study.

II. RELIABILITY ALLOCATION METHOD FREQUENTLY USED

Reliability allocation is the process of decomposing reliability indicators of the top-level system to the underlying unit step by step in the product design phase. In the course of the reliability allocation process, various methods should follow some common basic principles [12,13], such as:

 ISBN: 978-1-61208-126-7

TABLE I. FREQUENTLY-USED RELIABILITY ALLOCATION METHODS AND CHARACTERISTICS

Allocation method	Applicable stage	Applicable Scope	Applicable Conditions	Advantages	Disadvantages
Equal allocation method	demonstration stage	Basic reliability allocation and mission reliability allocation	When the product definition is not very clear	Simple calculation and convenient for use	Not giving considerations to the actual differences between various subsystems
Combination in proportion technique	Preliminary design stage	Conventional combination in proportion technique applies only to basic reliability allocation	With a similar physical model available for reference, requiring that model shall have certain data basis	Using the original actual user-defined data	High dependence on the reference model
Score allocation method	Preliminary design stage and detailed design stage	Basic reliability allocation and mission reliability allocation of series systems	With scarce reference data, with certain quality basis of technical personnel	The initiative and engineering experience of personnel can be given into full play, and score results has a certain convergence	Only scoring a few aspects of indicators, the results cover incomplete information

- Allocate lower reliability value to higher-complexity subsystems and equipment;
- Allocate lower reliability value to equipment with immature technology;
- Allocate lower reliability value to the equipment working in difficult environmental conditions;
- Allocate lower reliability value to products in service for long term;
- Allocate relatively higher reliability value to products with higher importance;
- Allocate relatively higher reliability value to products difficult to repair and replace;
- On-shelf products that have reliability value or systems that have been used sophisticatedly will not be allocated with new reliability value in design phrase, while it is required to remove the reliability value of these units from the total index then allocate the remnant reliability value to other systems.

Mainly due to different considerations and different premise of data application, different reliability allocation methods have different accuracy of their results. Frequently-used reliability allocation methods and their characteristics [14,15] are as shown in Table 1.

Although these above methods have corresponding advantages in different aspects, but there are still limitations in practical engineering applications:

1) Logical complexity of system functions. Owing to the interconnect, backup, timing and other related relations between functions of various system component units, the underlying units could not meet the fault independence assumption. Therefore the existing series, parallel, by-pass, bridging and voting and other reliability models could not describe the whole system mission reliability;

2) System reliability could not be calculated by the existing formula due to system and component unit life expectation does not obey the exponential distribution. Analytic deduction of other types of distributions is too complicated;

3) Each result of reliability allocation needs to be verified, the reliability allocation work can be put to an end if only the requirements are satisfied. But this process is too long for complex systems which require iterative calculation [16].

III. SIMULATION-BASED MISSION RELIABILITY ALLOCATION OF COMPLEX SYSTEMS

A. Allocation ideas

Reliability allocation and reliability prediction are two inverse processes, while the latter is used to validate the reasonableness of the results of reliability allocation. Simulation-based reliability allocation method of complex systems just uses this idea, i.e. in the case of a known initial value of reliability allocation, using the prediction method to obtain the simulation-based system reliability value and determine whether the initial requirements of allocation can be meet; if not so, it is required to regulate the values obtained in the current allocation according to the equipment importance or system weaknesses and other factors obtained from the simulation analysis, and then conduct prediction again until system requirement can be satisfied. This method does not require establishing the mission reliability model of the system while directly employs the criterion for system failure to building simulation models for the complex functional logic relationship and timing relationship between systems, which is applicable to any allocation types including exponential allocation. As a result of simulation-based calculations, after the initial establishment of system simulation models, we can conduct a lot of iterations and also weigh different allocation schemes without adding additional work. Therefore, it is suitable for the use in reliability allocation of complex systems in the engineering project.

B. Allocation process

Simulation-based reliability allocation mainly complies with the following steps:

First, make clear the definition of the system mission and establish fault criteria, which is the basis of the entire simulation. Simulation model can be established accurately

 ISBN: 978-1-61208-126-7

only different missions have been given a clear definition of fault criteria. Therefore, the definition of fault criteria must be clear enough: what unit works on what time phase, what is the influence of unit fault on the mission—the system fault or functional degradation, as well as what kind of combination and what kind of timing of each unit fault will lead to system fault.

Secondly, build simulation models based on the system functional logical relationship and fault criteria.

Thirdly, obtain the initial allocation value of the underlying unit by using conventional reliability allocation method. At this stage, if there is a known basic reliability requirements of the system, the result of basic reliability allocation can be used as the initial value assigned, so that the mission reliability allocation can be conducted under the premise of guaranteeing the basic reliability requirements so as to reduce the number of iterations.

Then, use the initial allocation value as an input of simulation model to run the simulation, and thereby compute the reliability of top system.

Finally, determine whether the top system reliability can meet the requirements; if not so, it is required to re-regulate reliability values of the underlying unit and then continue the simulation by iteration method until the requirement can be satisfied. In the process of values regulation, take importance factors into account so as to efficiently and rationally complete reliability allocation process [17].

Using the device as the underlying unit, the specific simulation algorithm is as shown in Figure 1.

(1) Build simulation models;
(2) Read the database to obtain the initial value of Mean Time Between Critical Fault (MTBCF) of each device, the number of device and other input parameters;
(3) Give simulation times, the maximum simulation time and time intervals between statistics;
(4) Running the simulation. Initializing, making all devices be from free faults;
(5) Generate random numbers; conduct random sampling on the life expectation of n device by using Monte Carlo method according to life expectation of device and MTBCF.

In the simulation running at J time, the fault time of the i^{th} device is:

$$t_{ij} = F_i^{-1}(\eta_{ij}) \quad (1)$$

In the formula, $F_i(t_{ij})$ is the fault distribution function of the i^{th} device; η_{ij} refers to the random number of random sampling of the i^{th} device at J time of simulation.

(6) Sequence by the order of the time of device fault occurring, let the fault first occur to the device with short fault time;
(7) According to the established simulation model, using traversal search method, sequentially determine whether the device fault results in system fault until the system fault occurs, identify system fault time t_j ;
(8) Record current fault time of the system and the serial number of devices to which faults have occurred;
(9) If the current simulation times are greater than the set simulation times, then go to step 10; otherwise, go to step 4;
(10) Using the interval statistical method to carry out system fault distribution statistic analysis for all data obtained by simulation to get the system MTBCF

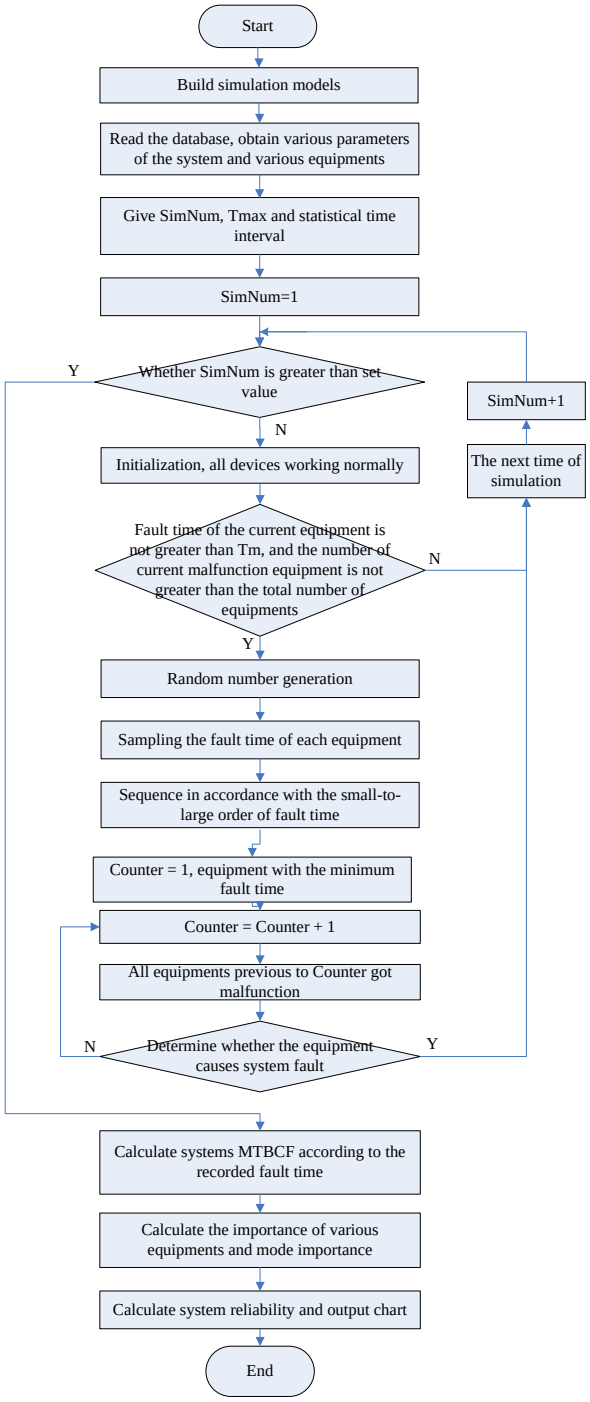


Figure 1. simulation algorithm

value, reliability, device importance and the mode importance. Importance reflects the importance degree of device in system. Mode importance is used to identify the weakness in system.

As the system fault time values are generated by random sampling, it is necessary to conduct statistical analysis for these values after N times of simulation. Suppose the maximum working hours of the system is $T_{\max}$ and equally divide it into several intervals (setting as m), then each time interval ΔT is:

$$\Delta T = \frac{T_{\max}}{m} \tag{2}$$

Relevant reliability indicators can be obtained by placing the system fault time of each simulation on its corresponding time interval, such as MTBCF and fault distribution function are respectively as follows:

$$MTBCF = \sum_{r=1}^{m} [t_r \cdot p_s(t_r)] \tag{3}$$

$$F_s(t_r) = \frac{m_r}{N} = \sum_{r=1}^{m} p_s(t_r) \tag{4}$$

In the formula, t_r refers to the biggest moment in the time interval r, $p_s(t_r)$ refers to the probability of placement in the time interval r, i.e. $p_s(t_r) = \Delta m_r / N$. In which, the importance of device is:

$$W(Zi) = \frac{\text{times of system in fault caused by equipment Zi in fault}}{\text{times of equipment Zi in fault}} \tag{5}$$

Mode importance of the device is:

$$W_N(Zi) = \frac{\text{times of system in fault caused by equipment Zi in fault}}{\text{times of system in fault}} \tag{6}$$

IV. APPLICATION OF WARSHIP SYSTEMS CASES

Take Warship Integrated Platform Management System as an example to do the reliability allocation.

A. System description and mission failure criterion

1) System Description

Warship Integrated Platform Management System is a very important system in warships, which is mainly used for warship implementing real-time monitoring, control and management over main systems (equipment) on the platform. This system contains eight sub-systems including 16 types and 82 equipments in total. Among them, one is a ship-borne computer which does not need to be allocated new reliability value and four equipments can be used as backup units for backup related equipments in all sub-systems. Thus between the different sub-system does not conform to the faults independent assumptions. There are series, parallel, n out of r and some fault correlation between interior equipments of different sub-systems. Therefore, it is very difficult to establish the system reliability models using conventional methods. We can describe the system in simulation method presented in 3.2.

2) Mission reliability requirements

MTBCF of Integrated Platform Management System is required to be 2000 hours.

3) Mission fault criterion

Warship working process will have different mission profiles. Under the corresponding different mission profiles, warship integrated platform management system has different working condition and fault criteria. For a typical mission profile, each sub-system will respectively achieve different functions. As a result, if any sub-system was in fault, the entire system functions could not be achieved. In the interior of sub-systems, some equipments fault will not lead to complete loss of functions of sub-systems but lead to functional degradation; fault relation exists between some equipments.

Taking No. 6 sub-system as an example, which includes No.75~No.79 equipment, No.17, No.51, No.52 and No.80 are four general-purpose backup units mentioned above, they are in other sub-systems and can backup No.75 equipment. Then, the fault criterion of No.6 subsystem is as follows: one of No.75~No.79 equipment is in fault, or No.75, No.17, No.51, No.52 and No.80 all get malfunctions.

B. Establish the simulation model of mission reliability

System mission reliability simulation model could be established according to the mission criteria. Functional faults correlations in No. 2 and No.6 subsystems are as shown in Figure 2 and Figure 3. In the graphics, the figure in box respectively represents equipment serial number. No. 51, No. 52 and No. 80 equipment do not belong to No. 2 sub-systems, but they can achieve backup for No. 17 equipment in the sub-system, so we can consider that they follow 1out of 4 relationship. After identifying functional fault correlations between various sub-systems, by following the series relationship between sub-systems, we can create reliability simulation model of the entire system, that is, the combination of functional faults and timing between what equipment can lead to system fault. System simulation model could be established according to the functional faults correlations in system, and then simulated program could be compiled according to the algorithm shown in figure 1. The partial codes are as follows.

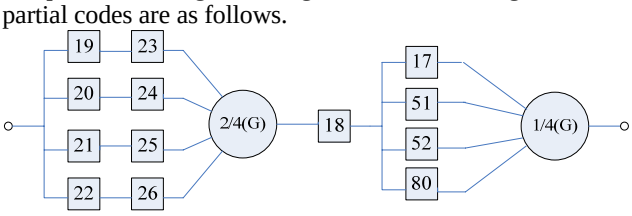


Figure 2. Functional faults correlation in No. 2 sub-system

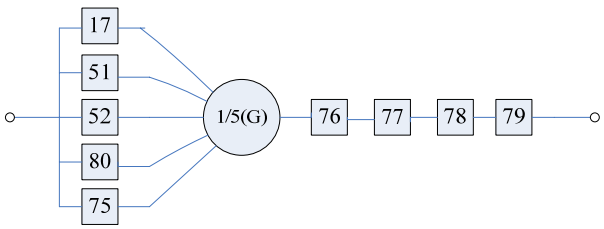


Figure 3. Functional faults correlation in No. 6 sub-system

 ISBN: 978-1-61208-126-7

```
…
(Get fault time and Sequence in accordance with the small-to-large order)
For j = 1 To 81
   E(j) = False   'Initialization, all equipments working normally
Next
For j = 1 To 81
    E(TTF(j).SerialNumber) = True   'the equipments are in fault according to the fault time order
     (Determine whether the equipment fault cause system fault)
    …
    S-S6_status = E(76) Or E(77) Or E(78) Or E(79) Or (E(17) And E(51) And E(52) And E(80) And E(75))
    'No.6 subsystem status
    …
    System_status = S-S1_status Or S-S2_status Or S-S3_status Or S-S4_status Or S-S5_status Or S-S6_status Or S-S7_status Or S-S8_status   'The whole system status
    …
     (fault data statistics and compute MTBCF, importance and mode importance)
    …
    Next j
    …
```

C. Obtain the initial value by series model allocation

MTBCF of No. 82 equipment can be up to 100,000 hours, after removing No. 82 equipment with the known fixed value from the system reliability indicator, the reliability allocation can be carried out in series model using score allocation method, and the result shall be taken as the initial value of simulation-based mission reliability allocation. Considering the complexity and working hours, the result of reliability allocation in series model is as shown in Table 3. The same type of equipment could be allocated with the same value.

TABLE II. SEQUENCE OF EQUIPMENT IMPORTANCE

Sequence number	Equipment Number	Importance	Mode importance
1	81	1.0000	0.0706
2	82	1.0000	0.0697
3	49	1.0000	0.0559
4	7	1.0000	0.0520
5	9	1.0000	0.0508
6	11	1.0000	0.0481
7	13	1.0000	0.0469
8	12	1.0000	0.0461
9	10	1.0000	0.0460
10	8	1.0000	0.0455
11	78	1.0000	0.0449
12	76	1.0000	0.0439
13	14	1.0000	0.0437
14	77	1.0000	0.0420
15	79	1.0000	0.0412
16	18	1.0000	0.0338
17	16	1.0000	0.0323
18	50	1.0000	0.0316
19	15	1.0000	0.0306
20	48	1.0000	0.0261
21	47	1.0000	0.0223
22	41	0.1490	0.0062
23	42	0.1313	0.0052
…	…		…

D. Run the simulation to obtain the final value of reliability allocation

Substitute initial reliability allocation values of various equipments into simulation model and run the simulation, and then obtain the system MTBCF up to 6630.548 hours, which is far greater than 2,000 hours as required by the system, so it is required to appropriately reduce the reliability index of some equipment. The main principles of adjustment: equipment with high importance, weakness links in systems shall reserve relatively higher reliability index; equipment assigned with unduly high reliability index probably could not be achieved in the project, so the reliability index should be appropriately reduced based on the actual situation.

According to the sequence of equipment importance and vulnerability in the system obtained in the first round of simulation (see Table 2), based on the practical application of actual engineering projects, we can regulate MTBCF values of some equipment and obtain system MTBCF 2247.826 hours after several rounds of iterations, so the current initial value can meet the requirements of mission reliability allocation, then reliability allocation work can be ended. The final allocation results are as shown in Table 4.

V. CONCLUSION

Using existing common methods to conduct system reliability allocation will be constrained in different aspects to varying degrees; especially in the case of complex system mission reliability allocation, some methods even could not solve the practical problems. However, in the case of using digital simulation ideas to conduct complex system mission reliability allocation, as long as the system mission and functions are identified, we can establish a system simulation model and further conveniently conduct verification on the results of complex system mission reliability allocation by using the initial allocation index of various underlying units, and meanwhile acquire the information of importance of various units, mission reliability and MTBCF of the whole system and various subsystems. In addition, this method can guarantee the correctness of multiple iterations testing and applies to application in practical engineering.

Using simulation-based complex system reliability allocation method is required to established corresponding simulation models for different application objects, which requires that engineering staff shall not only have a deep understanding of system functions prior to the establishment of simulation model, but also certain programming capabilities so as to ensure the correctness of establishing logic simulation models. Therefore, the research on general-purpose modeling simulation of complex systems can be further carried out in order to better apply to engineering projects.

ACKNOWLEDGMENT

The authors would like to thank Tu Qingci of BUAA for her review and thank Jia Xiaonan for her proof. The authors

 ISBN: 978-1-61208-126-7

are also grateful to the anonymous reviewers whose comments helped improve the presentation of the paper.

REFERENCES

[1] N. Schneidewind, "Allocation and analysis of reliability: multiple levels: system, subsystem, and module", Innovations System Software Engineering, 2006, vol. 2, pp. 121-136.

[2] J.E. Falk, N.D. Singpurwalla, and Y.Y. Vladimirsky, "Reliability allocation for networks and systems", Society for Industrial and Appllied Mathematics, 2006, vol. 48, pp. 43-65.

[3] Y.L. Wang, F.Y. Li, and J.F. Li, "Hybrid Method of Reliability Distribution for the Life Cycle of Product", Modular Machine Tool & Automatic Marnufacturing Technique, vol. 12, 2006, pp. 109-112.

[4] H.B. Zhang and Z.X. Jia, "Complex System Reliability Allocation based on Fuzzy Decision Method", 2009 International Workshop on Intelligent Systems and Applications.

[5] Y.K. GU and K.Q. Huang, "Fuzzy Reliability Allocation Method for Engine based on the Experts Knowledge", Journal of Aerospace Power, vol. 24, 2009, pp. 1143-1149.

[6] G.L. Lee, H. J. Lin, T.W. Yu, and C.C. Ma, etal. "Optimal Allocation for Improving System Reliability Using AHP", 2008 IEEE International Conference on Sustainable Energy Technologies, pp. 159-163.

[7] Y.C. Chang, K.H. Chang, and C.S. Liaw, "Innovative reliability allocation using the maximal entropy ordered weighted averaging method", Computers & Industrial Engineering, vol. 57, 2009, pp. 1274-1281.

[8] Z.W. Guo and G.B. Bai, "Reliability Allocation Using Analytical Target Cascading Method", Proceedings 2009 IEEE 16th International Conference on Industrial Engineering and Engineering Management, pp. 1195-1199.

[9] J. Ali, "Calculation of reliability allocation factor using sensitivity evaluation method", Proceedings of 2009 8th International Conference on Reliability, Maintainability and Safety, 2009, pp. 83-86.

[10] W.M. Yang and Y.X. Sheng, "System Reliability Digital Simulation", Beijing Aeronautics and Astronautics Press, 1990.

[11] X.H. Liu, G.H. Xu, C.C. Ma and S.F. Li, "Reliability Allocation and Optimization of Engine System by Using Genetic Algorithm and Monte Carlo Method", Journal of Beijing Institute of Technology, vol.16, 2007, pp. 268-273.

[12] S.K. Zeng, T.D. Zhao, and J.G. Zhang, etal. "System Reliability Design and Analysis", Beijing Aeronautics and Astronautics Press , 2001.

[13] H.M. Sun, "Some Reliability Methods and Application for Complex Product System", Qingdao University, 2007.

[14] X.H. Li, "Reliability Allocation Method of Integrated factors in a Series System", Journal of Baoji University of Arts and Sciences(Natural Science Edition), vol. 27, 2007, pp. 70-72.

[15] A. Yalaoui, C.B. Chu, and E. Chatelet, "Reliability Allocation Problem in a Series-parallel System", Reliability Engineering and System Safety, vol. 90, 2005, pp. 55-61.

[16] Y.J. Liu and M. Liu, "Research on Reliability Allocation of Complicated System", Journal of Projectiles.Rockets.Missiles and Guidance, vol. 24, 2004, pp. 77-79.

[17] J. Zhang and Y.C. Lei, "A method for allocating series system reliability based on analytic hierarchy process", Machinery Design & Manufacture, 2000, pp. 1-6.

TABLE III. INITIAL ALLOCATION RESULT

type	n	r_{i1}	r_{i2}	r_{i3}	r_{i4}	W_i	n* W_i	$C_i = wi / w$	$MTBCFi$
A	16	10	1	7	1	70	1120	0.0163284	124985.4
B	21	6	1	10	1	60	1260	0.0139958	145816.3
C	2	7	1	6	1	42	84	0.0097971	208309
D	4	6	1	8	1	48	192	0.0111966	182270.4
E	4	6	1	8	1	48	192	0.0111966	182270.4
F	1	6	1	8	1	48	48	0.0111966	182270.4
G	3	6	1	10	1	60	180	0.0139958	145816.3
H	1	10	1	8	1	80	80	0.0186611	109362.2
I	1	4	1	7	1	28	28	0.0065314	312463.6
J	1	7	1	5	1	35	35	0.0081642	249970.9
K	1	8	1	9	1	72	72	0.016795	121513.6
L	1	7	1	6	1	42	42	0.0097971	208309
M	14	6	1	6	1	36	504	0.0083975	243027.2
N	10	6	1	6	1	36	360	0.0083975	243027.2
O	1	10	1	9	1	90	90	0.0209937	97210.88
Sum	81						W =4287		

Note: n refers to the number of equipment of the same type; refers to the complexity; technical level; working hours; environmental conditions; score of various devices; coefficient of scoring for each unit; score of the system.

TABLE IV. FINAL ALLOCATION RESULT

type	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
MTBCF (ten thousand hours)	4	4.5	4.5	3	3	4.5	3.5	3	4.5	4.5	4.5	4.5	3	3	4.5

 ISBN: 978-1-61208-126-7

Simulation and Characterization of WLAN Indoor Channels at 60 GHz

Iskandar Iskandar and Dwi Harinitha
School of Electrical Engineering and Informatics, Bandung Institute of Technology (ITB)
Jalan Ganesha no.10 Bandung 40132 INDONESIA
Email: iskandar@ltrgm.ee.itb.ac.id, owie_dh@students.itb.ac.id

***Abstract*—This paper simulates and characterizes the 60 GHz band wireless channels for typical indoor. The Radio wave Propagation Simulator (RPS) is used to simulate the model characteristics. Simulations performed on the three rooms at Telematics Laboratory of Bandung Institute of Technology (ITB) those are Residency room, Computer Laboratory (size 1013 cm x 589.5 cm x 289.8 cm), and Undergraduate/Final Project room (size 900 cm x 1013 cm x 289.8 cm). The rooms were built with concrete walls in the left and right side with a thickness 14 cm, 0.5 cm for ceiling thickness, and the combination of walls and glasses at the front and backside of the rooms. The model characteristics were measured based on three classifications of distance between transmitter and receiver, i.e., CM1 (0 - 4 m), CM2 (4 - 10 m), and CM3 (> 10 m). We then obtain the model characteristics in terms of RDS value at 0.26 to 7.74 ns for the receivers at Residency room, 1.02 to 8.76 ns at Computer Lab and 2.07 to 15.65 ns at Final Project room. The system performances shown the coherence bandwidth is 1.2778 to 76.9492 MHz for frequency correlation above 90% and therefore the maximum bit rate that can be achieved with 1.25 spectrum efficiency is 60 Mbps.**

***Keywords*—60 GHz, RPS, rms delay spread, coherence bandwidth.**

I. INTRODUCTION

Wireless network is no longer an exclusive technology. The utilization of this wireless network can served as an alternative for computer network communication system whether in large scale or small scale that can be choose by someone who preferred practical non-wire-communication. This service can be found for private area (WPAN-Wireless Personal Area Network), local area (WLAN-Wireless Local Area Network), and public area (WMAN-Wireless Metropolitan Area Network), i.e., office buildings, school or campus, supermarket and residential home.

In wireless network realization, problems will occur that consist of data rates, coverage area, network size and power consumption, however, this can be fixed using WPAN (Bluetooth) or WLAN (Wi-Fi) which has different data rates. WPAN has lower data rate characteristics compared to WLAN, therefore WLAN application is more developed. Bandwidth which being used at present is 2.4 GHz (ISM-Industrial, Scientific and Medical), unfortunately, this bandwidth is jam-packed with many applications that runs in short and long distance transmission, while the demand on higher data rates are increasing. Therefore, frequency around 60 GHz (millimeter waveband), which is unlicensed frequency, is becoming more interesting. The new technologies are designed to develop the gigabyte wireless connection although there are many challenges ahead [5][7].

The most important issue in this frequency applications are the channel modeling and characteristics. Reference [2] shows measure and characterization of wireless channel using S-V Model. However, this analytical research is not sufficient enough, not to mention the frequency integration of 2.4/5 GHz to 60 GHz [7].

Concerned with the model that being used and to respond with the future challenges, this research are based on the wireless indoor channel characterization at 60 GHz through computer simulation so that we can identify the channel behavior.

To limit the scope of this work, we focus our research on the delay spread value including mean excess and rms delay to obtain the main wireless indoor parameters. The rest of the paper is organized as follows. Section 2 reviews the wireless channel, propagation and channel modeling. Section 3 describes the simulation procedure and the model. Section 4 shows simulation results in terms of the main parameters. We draw conclusions in Section 5.

II. CHANNEL MODELING

In wireless communication systems, the propagation characteristics are having influence on the system design, both for outdoor or indoor. For 60 GHz, characteristics indoor system have significant differences (very unique) compared to outdoor system, are mainly caused by obstacles, i.e., walls, floor, ceilings and many objects which cause the radio signals will experience the reflection, diffraction and scattering effect. In addition to the basic mechanisms of propagation that associated with the fading and multipath, there is also attenuation due to absorption of oxygen in the amount of 10-15 dB/km which makes the frequency is not suitable for long distance (> 2 km). Therefore, applications for 60 GHz dedicated to indoor, so that characteristic of indoor communication systems should be considered especially the delay spread as multipath effects.

To predict the delay spread, modeling based on deterministic propagation model that is ray tracing model with a computational model (optical ray technique). Propagation described by the wave propagation of different signals from transmitter to receiver antenna with respect to reflection, diffraction, and scattering objects in buildings and other obstructions. For computational process, is based

 ISBN: 978-1-61208-126-7

on the general theory of diffraction, that is the Universal Geometrical Theory of Diffraction (UTD). In this computational process, is required much time, especially in determining where the relevant point between transmitter and receiver. However, the accuracy level is higher because of obstructions or objects are taken into account during the propagation. Figure 1 shows a computing propagation model with the ray tracing algorithm.

III. SIMULATION MODEL

The simulation will be performed using RPS software to obtain delay spread value within simulation areas wherein the 3D database of that area will be required. In this work we must considering the space area, shape, size and position of every objects within the area, included the material of all objects. The simulation area in Telematics Lab at ITB, that is S2/S3 Residency Room, Computer Lab, and Final Project Room.

Based on the database mentioned earlier, we create the 3D area image using IDE and arrange the system configuration both on transmitter and receiver. Transmitter antenna using horn antenna with 20 dB gain positioned at the corner of the area with 2.5 m height and transmit power of 10 dBm. At the receiver we employ dipole antenna with different heights and positioned in entire of area correspond to the data that will be taken which based on the distance between transmitter and receiver. We divide the distance into three groups, i.e., CM1 (0 – 4 m), CM2 (4 – 10 m) and CM3 (more than 10 m). In Residency Room we had 5 receivers for CM1 while for CM2 and CM3 we put 6 and 8 receiver, respectively. For the Computer Lab and Final Project Room, the receivers for CM1, CM2, and CM3 are (6, 8, 4) and (2, 8, 6), respectively. After configuring the systems we start the simulation using 3D ray tracing algorithm with total two reflection and penetration each and one diffraction. Fig. 2 shows the simulation area and system configuration.

We evaluate the system performance (based on the simulation results of area modeling) in terms of BER vs Eb/No which correspond to multipath fading condition that is flat and frequency selective fading. We use BPSK modulation with Rayleigh channel fading model and spectrum efficiency of 1.25 where bitrate as an analysis parameter.

IV. SIMULATION RESULTS

In this section, we analyze the data and the simulation results. Our target is the model characteristics and system performance (BER vs Eb/No) based on bit rate capacity (R_b).

For CM1 condition at Residence Room, we found that the RDS value for each receiver with minimum of 0.26 ns at Rx1 and maximum of 1.85 ns at Rx2. RDS value based on range (on average at CM1) is 1.18 ns. The color differences show the different experienced delay. In Fig. 3(a), the blue color indicates that the delay period is quite small because the distance between transmitter and receiver not to remote and also have not too many reflector objects. Ray propagation in 2D and 3D can be seen in Fig. 3(b) and 3(c).

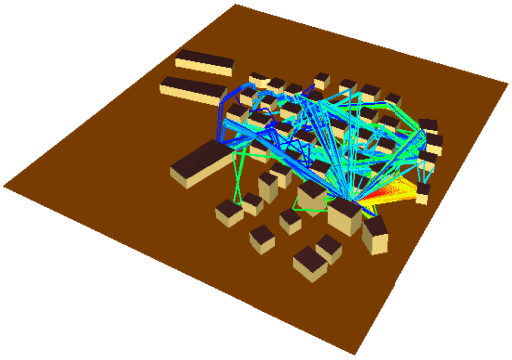


Figure 1. *Ray tracing.*

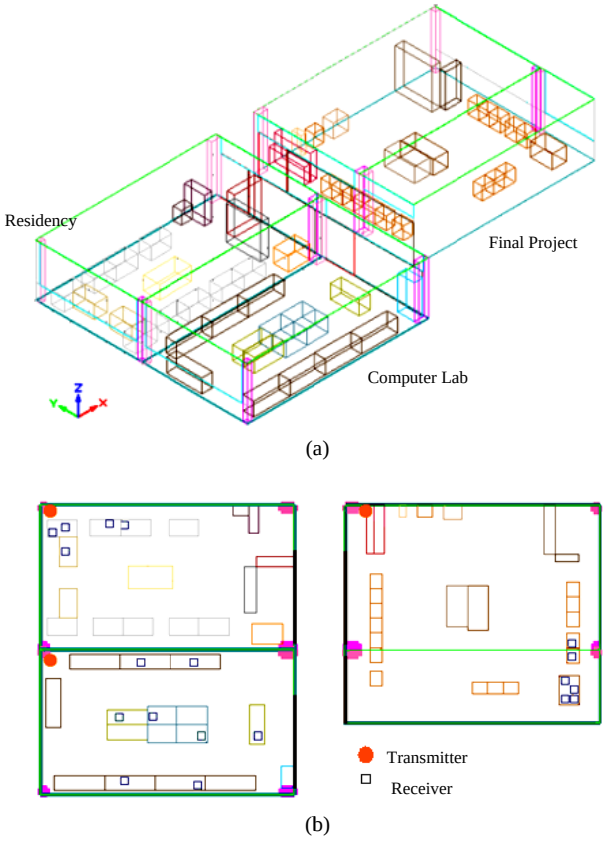


Figure 2. Simulation area : (a) 3D *database* and (b) system configuration.

Fig. 4 shows CM2 condition. The results indicate that if we increase the distance and the receiver for different position, it will also increase the reflection resulting in longer time for ray to arrive at the receiver. This can be shown from the color variation (Fig. 4(a)) that indicates larger delay compared to that in CM1. The minimum RDS for this condition is 1.94 ns at Rx2 and maximum 7.74 ns at Rx5. Averaging RDS is 4.05 ns.

Simulation results for CM3 can be seen in Fig.5. The minimum RDS is 1.10 ns at Rx1 and maximum 2.11 ns at Rx2 with 1.52 ns on average. The RDS value is quite small although the distance between transmitter and receiver are quite remote. This is caused by the fact that the longer distance the smaller the signal power is therefore the ability to reflect also reduced. Table 1 shown the complete results of this first category.

 ISBN: 978-1-61208-126-7

The measurements based on the receiver distance and position at identical area, for CM1 at Residence Room, position of Rx1 near with the transmitter and having 1.85 ns RDS while Rx4 at the farthest position having a 1.79 ns RDS. The minimum RDS at Rx3 with 1.03 ns and maximum at Rx4 with 1.85 ns.

We can see that at the Computer Lab, the min RDS is 0.65ns at Rx2 and maximum at Rx3 with 3.23ns. It indicates that although the receiver is located in equal position and distance, we still obtain different RDS value. This shown that distance and position is not a suitable parameters to obtain the maximum and minimum of RDS value. The RDS value for this category can be seen at Table 2.

We also perform the simulation at 2.4 GHz frequency. The most significant difference is in terms of coverage and ability to mitigate the reflection, diffraction and scattering. Fig. 6 and Table 3 shown the differences of the two frequencies.

For system performance we considered only the maximum and minimum RDS value at Residency Room. From Fig. 7 with flat condition, the BER performance of 10^{-3} is feasible up to 15 Mbps bitrate and move toward selective condition after 60 Mbps (BER > 10^{-3}). Theoritically this can be obtain through equation (1) for frequency correlation function above 90%.

At RDS of 0.26 ns, the coherence bandwidth with 0.9 frequency correlation is 76.92 MHz. The channel bandwidth must be smaller than coherence bandwidth in order to achieve flat fading condition. We choose 76 MHz so that the maximum bitrate is 60 MBps under flat fading. However this maximum condition still influenced with another factors so that the maximum bitrate value still vary.

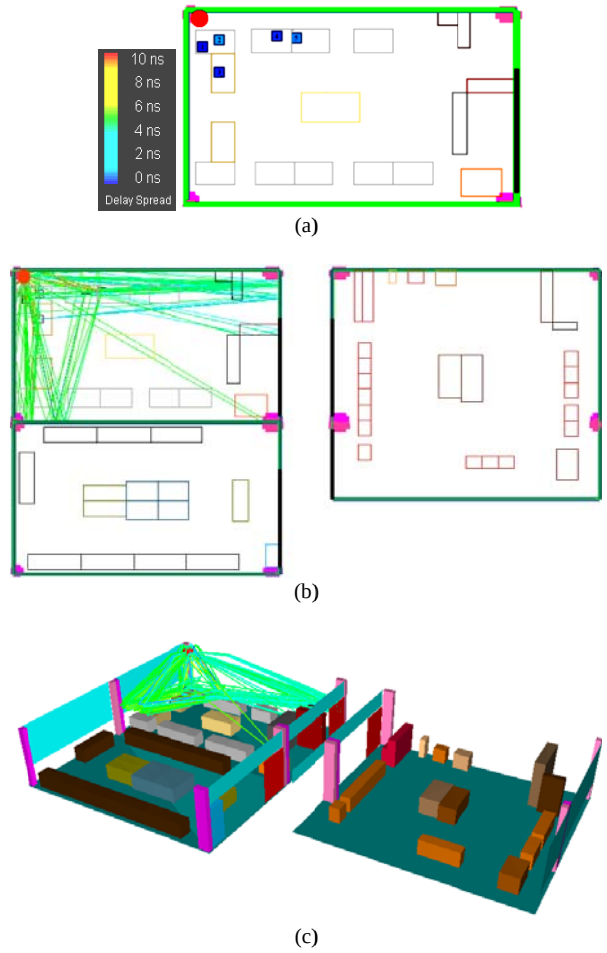


Figure 3. Simulation results for CM1 : (a) *delay*, (b) 2D, and (c) 3D.

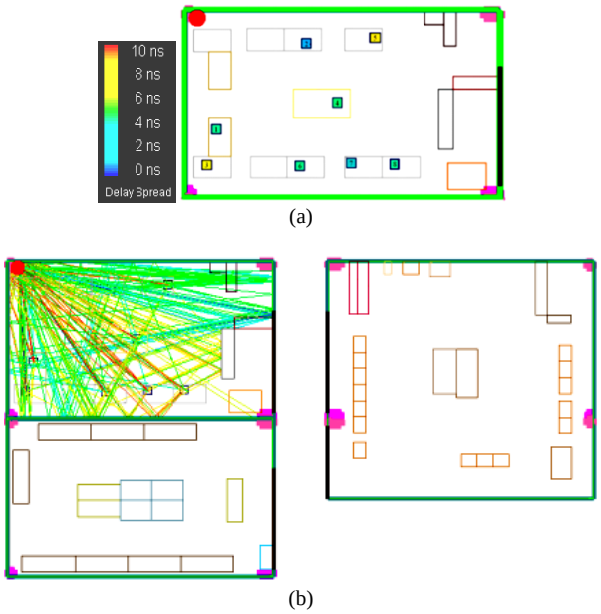
TABLE I. RMS DELAY SPREAD AT 60 GHZ.

Condition	Rx Position	Residency	
		RMS (ns)	average RMS
CM1	1	0.26	1.18
	2	1.85	
	3	1.04	
	4	1.14	
	5	1.59	
CM2	1	3.27	4.05
	2	1.94	
	3	7.51	
	4	3.46	
	5	7.74	
	6	3.28	
	7	2.07	
	8	3.13	
CM3	1	1.1	1.52
	2	2.11	
	3	1.23	
	4	1.58	
	5	1.22	
	6	1.85	

Condition	Rx Position	Computer Lab.	
		RMS (ns)	average RMS
CM1	1	1.43	3.39
	2	1.02	
	3	4.21	
	4	4.57	
	5	3.86	
	6	5.25	
CM2	1	1.12	4.01
	2	3.75	
	3	4.12	
	4	8.76	
	5	1.97	
	6	6.54	
	7	2.04	
	8	3.75	
CM3	1	2.76	2.90
	2	3.45	
	3	3.28	
	4	2.10	

Condition	Rx Position	Final Project	
		RMS (ns)	average RMS
CM1	1	4.04	6.17
	2	8.30	
CM2	1	7.57	7.31
	2	6.90	
	3	2.07	
	4	3.45	
	5	9.08	
	6	15.65	
	7	9.75	
	8	3.94	
CM3	1	5.41	4.01
	2	3.68	
	3	4.26	
	4	4.45	
	5	3.03	
	6	3.26	

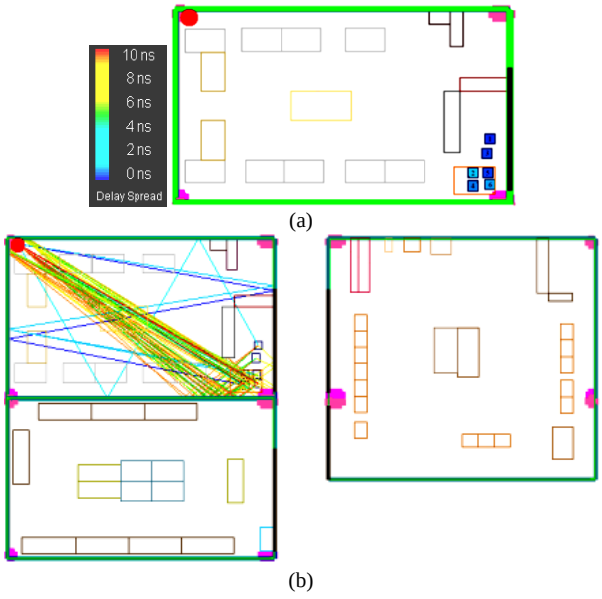
 ISBN: 978-1-61208-126-7



(a)

(b)

Figure 4. Simulation results for CM2 : (a) *delay* and (b) 2D.



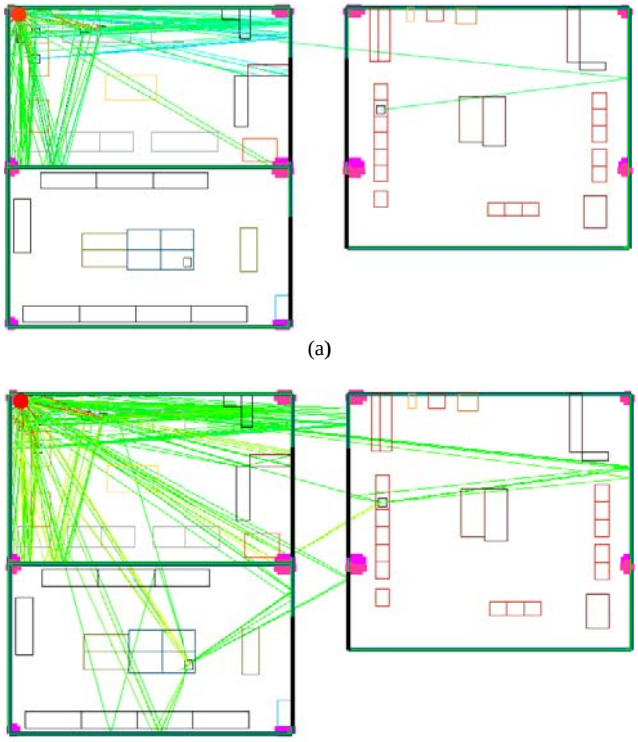
(a)

(b)

Figure 5. Simulation results for CM3 : (a) *delay* and (b) 2D.

TABLE II. RMS DELAY SPREAD AT IDENTICAL POSITION AND DISTANCE OF RECEIVER.

Condition	Rx Position	Residency		Computer Lab.	
		RMS (ns)	Average RMS	RMS (ns)	Average RMS
CM1	1	1.85	1,41	1.52	2,15
	2	1.04		0.65	
	3	1.03		3.23	
	4	1.79		3.19	
CM2	5	2.94	5,06	2.95	3,47
	6	7.47		3.45	
	7	11.1		3.83	
	8	8.37		6.75	
	9	2.91		3.86	
	10	3.24		1.57	
	11	2.56		1.69	
	12	2.97		2.91	
	13	4.04		4.22	
CM3	14	1.06	1,43	3.39	2,96
	15	1.58		2.69	
	16	1.22		3.04	
	17	1.85		2.74	



(a)

(b)

Figure 6. Comparison for 60 GHz and 2,4 GHz : (a) *ray* at 2.4 GHz and (b) *ray* at 60 GHz.

TABLE III. RMS DELAY SPREAD COMPARISON FOR 60 GHZ AND 2.4 GHZ.

Room	Condition	Rx Position	60 GHz		2.4 GHz	
			RMS (ns)	Average RMS	RMS (ns)	Average RMS
Residency	CM1	1	0.26	0.87	0.55	3.71
		2	1.86		2.18	
		3	1.05		2.09	
		4	1.24		1.95	
		5	1.70		2.10	
Computer	CM2	6	0.00		3.15	
Final Project		1	0.00		13.98	

With 1.85 ns RDS and same bitrate, the near-selective condition achieved at 10 Mbps whereas theoritically at 8 Mbps. This simulation result approaching the theoretical results however is not too noticeable at this condition since we do not perform the simulation for 8 Mbps bitrate. The max bitrate that can be achieved at min RDS is larger than maximum RDS and applied at all simulation results. The simulation results can be compared with theoretical estimation as shown in Table 4, while the compared results between 2.4 GHz and 60 GHz frequency can be seen in Table 5.

$$B_c = \frac{1}{50\,\sigma_s} \qquad (1)$$

From this work we can proves that the simulation results correspond to the estimation results and we can identify the correlation between RDS, coherence bandwidth and bit rate. With the greater RDS value, the coherence bandwidth will be increasingly narrow therefore the achieveable maximum bitrate will be smaller.

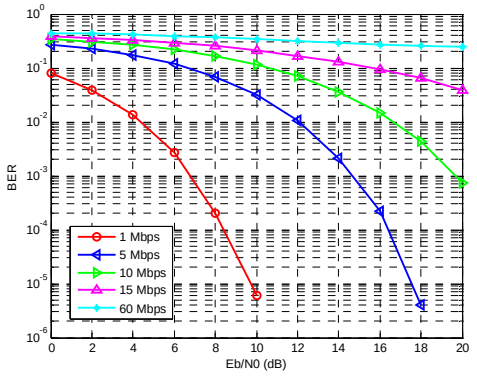
 ISBN: 978-1-61208-126-7

TABLE IV. COHERENCE BANDWIDTH AND BITRATE OBTAINED FROM MINIMUM AND MAXIMUM DELAY SPREAD.

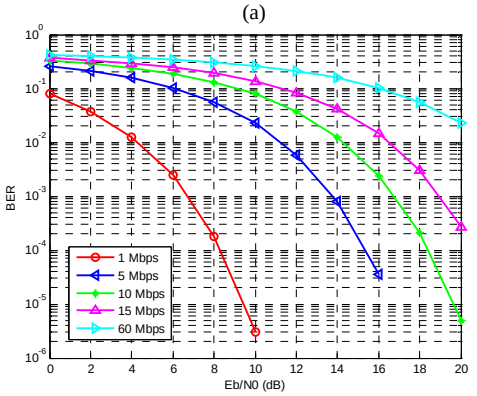
Room	Condition	RMS (ns)	Bc (MHz)	Rb (Mbps)
Residency	CM1	1.18	16.94915	12
	CM2	4.05	4.93827	3
	CM3	1.52	13.15789	10
Computer	CM1	3.39	5.89971	4
	CM2	4.01	4.98753	3
	CM3	2.90	6.89655	4
Final Project Room	CM1	6.17	3.24149	2
	CM2	7.31	2.73598	1
	CM3	4.01	4.98753	3

TABLE V. COHERENCE BANDWIDTH AND BITRATE AT 2.4 GHZ.

Room	Condition	RMS (ns)	Bc (MHz)	Rb (Mbps)
2.4 GHz	Min.	0.55	36.36364	28
	Max.	13.98	1.43062	1
	Area	3.71	5.39084	4
60 GHz	Min.	0.26	76.92308	60
	Max.	1.86	10.75269	8
	Area	0.87	22.98851	17



(a)



(b)

Figure 7. Performance for CM1 with RDS : (a) minimum 0.26 ns and (b) maximum 1.85 ns.

V. CONCLUSION AND FUTURE WORK

Based on simulation results and analysis at three simulation area we can conclude that the indoor channel characteristics are influenced by the objects and the object materials inside the area itself. The wavelength will affecting the reflection, if the frequency higher therefore the reflection effect will be larger. The model characteristics for 60 GHz, maximum RDS value achieved at 15.65 ns in Final Project room while minimum value achieved at 0.26 ns in Residency room. Using frequency correlation function above 90%, the coherence bandwidth is 1.2778 – 76.9492 MHz therefore the maximum bitrate that can be achieved with 1.25 spectrum efficiency is 60 Mbps.

This contribution is limited to an experimental ray-tracing simulation for indoor 60 GHz transmission. The measurement is required to validate the simulated channel model. Our future work will be focused to validate the model by using experimental measurement.

REFERENCES

[1] A. Deshmukh and S.K. Bodhe, "Characterization of Radio Propagation at 60 GHz Channel," *First Asian Himalayas International Conference*, pp. 1-8, 2009.

[2] C. Liu, E. Skafidas, and R.J. Evans, "Characterization of The 60 GHz Wireless Desktop Channel," *IEEE Transaction on Antenna and Propagation*, Vol. 55, no. 7, pp. 2129-2133, 2007.

[3] R. Felbecker, W. Keusgen, and M. Peter, "Ray-Tracing Simulations of the 60 GHz Incabin Radio Channel," *URSI : XXIX General Assembly of the International Union of Radio Science*, Chicago, IL, USA, 2008.

[4] A. Goldsmith, *Wireless Communications*. California : Cambridge University Press, 2002.

[5] N. Gou, R.C. Qiu, S.S. Mo, and K. Takahashi, "60-GHz Millimeter-Wave Radio : Principle, Technology, and New Results", *EURASIP Journal on Wireless Communications and Networking*, Vol. 2007, 8p, 2007.

[6] H. Nam, *Statistical Analysis of Indoor Radio Propagation Channel Characteristics at 60 GHz*, Project Report (EE381V), The University of Texas at Austin, 2003.

[7] M. Park, C. Cordeiro, E. Perahia, and L.L Yang, "Millimeter-Wave Multi-Gigabit WLAN : Challenges and Feasibility", *PIMRC, IEEE 19th International Symposium*, 2008, pp. 1 – 5.

[8] L. Rakotondrainibe, G. Zaharia, G.E. Zein, and Y. Lostanlen, *Indoor Channel Measurements and Communications System Design at 60 GHz*. URSI, Versi 1, 2008.

[9] T.S. Rappaport, *Wireless Communications : Principles and Practice*. 2nd Edition, New Jersey : Prentice Hall PFR, 2002.

[10] P.F.M. Smulders, 60 GHz Radio : Prospects and Future Directions,*SCVT*,2003,http://tte.ele.tue.nl/radio/publications/ECR%20pubs%202003/Smulders%2060GHz%20radio%20SCVT2003.pdf (accessed : 18 Feb 2009).

[11] P.F.M. Smulders, M.H.A.J. Herben, and J. George, "Application of Five-Sector Beam Antenna For 60 GHz Indoor Wireless Communications," *IET Journals*, Vol. 38, pp. 1054 – 1055, 2002.

[12] P.F.M. Smulders, C.F. Li, H. Yang, E.F.T. Martijn, and M.H.A.J. Herben, 60 GHz Indoor Radio Propagation – Comprison of Simulation and Measurement Results, 2004, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.8825&rep=rep1&type=pdf (accessed : 15 Jun 2009).

[13] M.T. Supriyadi, *Studi Perbandingan Model Propagasi Empirik dengan Menggunakan Simulasi Ray-Tracing di Lingkungan Urban Bergedung pada Sistem Komunikasi Fix BWA Standar IEEE 802.16Revd Wimax*, Final Project, Institut Teknologi Bandung, 2006.

[14] H. Yang, P.F.M. Smulders, and M.H.A.J. Herben, "Indoor Channel Measurements Analysis in the Frequency Bands 2 GHz and 60 GHz," *PIMRC , IEEE 16th International Symposium*, Vol. 1, 2005, pp. 579 – 583.

[15] H. Yang, P.F.M. Smulders, and M.H.A.J. Herben, Frequency Selective of 60-GHz LOS and NLOS Indoor Radio Channels, *IEEE 63rd Vehicular Technology Conference (VTC2006- Spring)*, Vol. 6, 2006, pp. 2727 – 2731.

[16] S. Widodo,Gi-Fi Jadi Standar Komunikasi *Wireless* Baru?, 2008, http://web.bisnis.com (accessed : 15 Jan 2009).

 ISBN: 978-1-61208-126-7

A Comparison of Receiver Strategies in STBC MIMO Systems in a Challenging Environment

Krzysztof Kosmowski
Radiocommunications Department
Military Communications Institute
Zegrze, Poland
e-mail: k.kosmowski@wil.waw.pl

Józef Pawelec
Radiocommunications Department
Military Communications Institute
Zegrze, Poland
e-mail: j.pawelec@wil.waw.pl

***Abstract*—The subject of this paper is a comparison of the different criteria and strategies of signal reception in space-time coding MIMO systems operating under dynamic flat fading and channel cross-correlation. The dynamic fading means that a Doppler shift is comparable with the transmission rate. The criterion of comparison is a mean bit error rate. The BER characteristics are obtained via Monte Carlo simulation. Many strategies and antennas configurations are analyzed, including space-time maximum likelihood, and zero forcing, as well as 2x1, 2x2 and 2x3 antenna setup. Simulation demonstrates that the receiver with 3 antennas and simple maximum likelihood decoder usually acts better than receiver with a fewer number of antennas and a more sophisticated reception strategy.**

Keywords- STBC MIMO; ML; ZF; ML-ST

I. INTRODUCTION

Potential performance benefits and a remarkable capacity promised by MIMO (multiple input – multiple output) systems attracted a lot of interest in the recent years. However, to keep this promise, many strict conditions have to be satisfied. This refers, inter alia, to the assumption of a quasi-static fading where the channel gains h(t)=h(t+T) [1] [2]. Under real conditions, the transfer function h(t) changes itself over the code word and this leads to interference, supposing the Maximum Likelihood (ML) criterion is used [7][8]. At the same time, the interference suppression methods, e.g. Zero Forcing (ZF) or Minimum Mean Square Error (MMSE) techniques are applied [4][5][6], as well as the QR decomposition [10] and the Maximum Likelihood Space-Time decoding (ML-ST) [8]. These strategies are usually examined in MISO (multiple input - single output) configuration and in the absence of a cross-correlation. Those approaches outperform linear combiner with the ML detector. However, they are more complex. The influence of the number of receiving antennas on this benefit was not taken into account, as well as the effect of cross-correlation.

In the present paper, several antennas architectures and popular detector schemes - ML, ML-ST, ZF, MMSE - are analyzed, both for channel cross-correlation and for dynamic flat fading. The comparison of performance is carried out on the basis of BER characteristics.

The rest of the paper is organized as follows. Section II describes the system model. Section III deals with interference suppression techniques. In Section IV, the Monte Carlo simulations are carried out and the obtained results are analyzed. Section V contains a concise conclusion.

II. SYSTEM MODEL

The considered system consists of two transmitting and one, two or three receiving antennas, Fig 1. All antennas are potentially mobile and omni-directional. It is assumed that the system exploits the space-time coding strategy [3]. In principle, the maximum likelihood decision rule is applied assuming that the interference suppression (IS) is performed first. The channel undergoes frequency flat and time-selective Rayleigh fading. The channel gains are identical Gaussian random variables with zero means and autocorrelation function $1/2\,E[h_i(t)h_i^*(t+\tau)]=\sigma^2 R(l)$ where $R(l)$ follows Jakes' model [12].

$$R(l) = J_0(2\pi F_d T_s l) \tag{1}$$

$J_0(\,)$ is the zero-order Bessel function of the first kind, F_d stands for maximum Doppler shift and T_s is the symbol duration time.

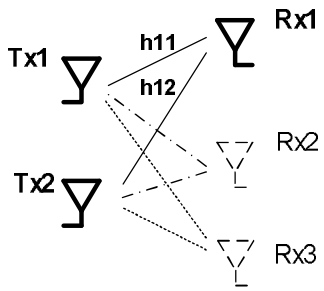


Figure 1. Considered MIMO System

We assume that the links between different antennas are identically distributed, however they can be correlated. The perfect estimation of the channel transfer functions *h(t)* is presupposed.

 ISBN: 978-1-61208-126-7

The considered scheme exploits the code word of Alamouti [3]. For the 2x2 MIMO system, received signals in antennas Rx1, Rx2 at some point of time t are r_{1t}, r_{2t} respectively, while at the next moment $t+T$ - r_{1T}, r_{2T}, where T – signal symbol duration and n_i are the independent and identically distributed (i.i.d.) complex Gaussian variables with zero mean and some variance N_0.

$$\begin{aligned} r_{1t} &= h_{11t}S_1 + h_{12t}S_2 + n_1 \\ r_{2t} &= h_{21t}S_1 + h_{22t}S_2 + n_2 \\ r_{1T} &= h_{12T}S_1^* - h_{11T}S_2^* + n_3 \\ r_{2T} &= h_{22T}S_1^* - h_{21T}S_2^* + n_4 \end{aligned} \tag{2}$$

The changes in the third and fourth row of (2) follow from the fact that signals S_1, S_2 in the $t+T$ period are conjugated and S_1 is multiplied by (-1). By means of manipulating the receive vector $r = [r_{1t}, r_{1T}, r_{2t}, r_{2T}]^T$ the matrix notation of (2) takes the form.

$$\begin{bmatrix} r_{1t} \\ r_{1T}^* \\ r_{2t} \\ r_{2T}^* \end{bmatrix} = \begin{bmatrix} h_{11t} & h_{12t} \\ h_{12T}^* & -h_{11T}^* \\ h_{21t} & h_{22t} \\ h_{22T}^* & -h_{21T}^* \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2^* \\ n_3 \\ n_4^* \end{bmatrix} \Leftrightarrow r = HS + n \tag{3}$$

The decoder proposed by Alamouti acts as follows [3]

$$\widetilde{S} = H^H r \Rightarrow \widetilde{S} = H^H HS + H^H n \tag{4}$$

where $\widetilde{S}$ is an estimate vector of transmitted symbols.
Ignoring in (4) the noise component $H^H n$ for a quasi-static channel, $h(t)=h(t+T)$, the estimates of signals are

$$\begin{bmatrix} \widetilde{S}_1 \\ \widetilde{S}_2 \end{bmatrix} = \begin{bmatrix} |h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2 & 0 \\ 0 & |h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2 \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \tag{5}$$

In real channels, however, the off-diagonal elements of $H^H H$ are no longer zero and this introduces an interference.

III. Interference Suppression

The main idea leading to interference suppression techniques is to built a matrix W to fulfill condition

$$WH = I \tag{6}$$

where I stands for diagonal matrix of real elements. In the ZF technique it takes a form [4]

$$W = (H^H H)^{-1} H^H \tag{7}$$

As a consequence, elements causing the interference are reduced to zero but the variance of system noise is at the same time increased, what obviously deteriorates system performance. The way to reduce this effect is to built a modified matrix (8) and minimize the noise via the MMSE technique [13].

$$W = (H^H H + \sigma^2 I)^{-1} H^H \tag{8}$$

where I -identity matrix and σ^2 - noise variance.
Most publications deal with systems with one receiving antenna. This assumption provides an opportunity to use a simple matrix inversion. However, for a few receiving antennas the problem of matrix inversion is more complex. Such a matrix is no longer the square one and the advanced pseudo-inversion has to be carried out [9]

$$A^+ = (A^T A)^{-1} A^T \tag{9}$$

Another way to perform such an operation is a singular value decomposition (SVD):

$$A = USV^T \Rightarrow A^+ = V(S^T S)^{-1} S^T U^T \tag{10}$$

Still another technique employs a space-time decoder, ML-ST [8]. It chooses the most probable sequence of the transmitted pair of symbols. The ML detector minimizes the product

$$\left| r - H\widetilde{x} \right|^2 \tag{11}$$

where $\widetilde{x}$ stands for pair of symbol, in BPSK case $\widetilde{x} \in \{(1\,1),(1,0),(0,1),(0,0)\}$.

IV. Simulation Results

In order to compare the different strategies the Monte Carlo simulation method was applied. As a measure of channel variations, the normalized fading rate $F_d T_S$ was used, where F_d stands for a maximum Doppler spread and T_S for a symbol duration time (at times, this parameter is called the normalized fading bandwidth, BT).

Three values of $F_d T_S$ were used: 0.01 for quasi-static fading, 0.05 for moderate fading and 0.1 for fast fading. The first results, Fig.2, refer to ZF and MMSE techniques. The benefit of using MMSE instead of ZF becomes more and more negligible as the number of receiving antennas exceeds one. For the 2x2 setup, both detectors provide comparable results even for fast fading, and for the 2x3 setup, the difference disappears. Such results were expected.

 ISBN: 978-1-61208-126-7

It is to be noted that the size of identity matrix I is dictated by the size of the $H^H H$ matrix in (8).

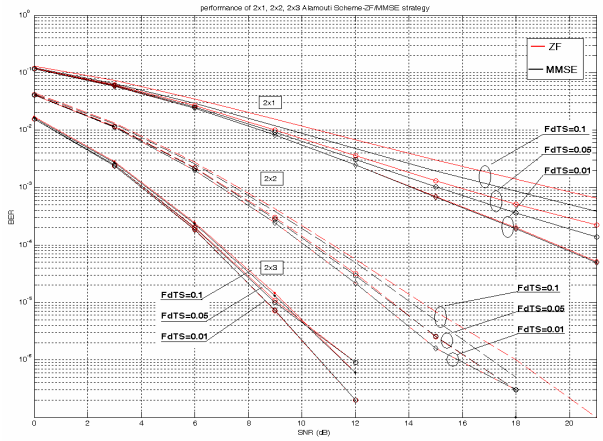


Figure 2. BER performance of ZF/MMSE strategies for 2x1, 2x2, 2x3 antennas setups and F_dT_S=0.01 - 0.1

For the Alamouti code, this operation always leads to a 2x2 square matrix irrespective of the number of receiving antennas, because there are always 2 transmitting antennas. The noise minimizing component $\sigma^2 I$ have the same value for all considered antennas configurations.

Thus, the benefit of using MMSE instead of ZF becomes smaller with the increase of the number of receiving antennas. One can see from Fig. 2 that the highest impact on the behavior of all characteristics evokes the number of receiving antennas – the higher the number, the lower the BER. The next factor is the type of detector: the ML gives the poorest results, ZF and ML-ST seem better and comparable to each other. The influence of a cross-correlation is temperate.

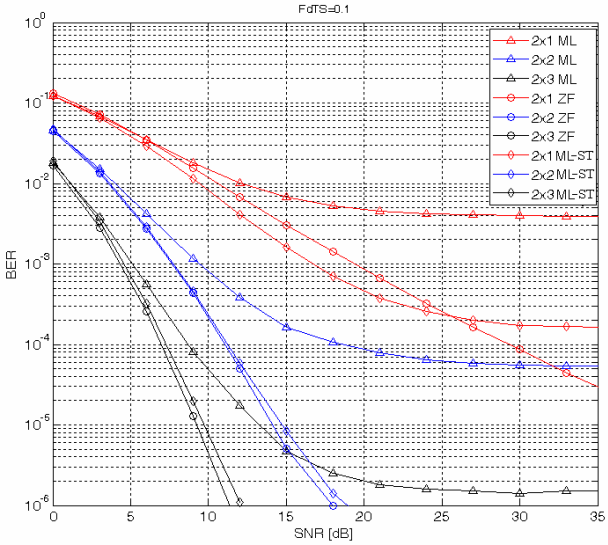


Figure 3. . BER performance of ML, ZF, ML-ST strategies for setups 2x1, 2x2, 2x3 and fast fading, negligible cross-correlation

However, it can be noticed that the ML detector in a 2x2 setup gives better performance than all considered detectors in a 2x1 configuration in a range of SNR from 0 to approx. 30dB.

Fig. 5 refers to strong channel cross-correlation and fast fading.

The a, b, c parts of the Fig. 5 refer to ML, ZF and ML-ST respectively. Colors denote the number of receiving antennas and

- solid lines – fast fading strong correlated case;
- dashed lines – fast fading uncorrelated case;
- dotted bold lines – moderate fading uncorrelated case as a reference level.

It can be seen that all considered strategies suffer from fading and cross-correlation. The greatest impact on the behavior of all characteristics again evokes the number of receiving antennas – the higher the number, the lower the BER. The next factor is the type of detector: the ML gives the weakest results, ZF and ML-ST seem better and comparable to each other. The influence of a cross-correlation is temperate.

The second group of results, Fig.3 and Fig. 4, refer to ML, ZF and ML-ST strategies (MMSE was neglected) for moderate fading and both uncorrelated channel and low channel cross-correlation. The elements of the correlation matrix were set randomly from the values of 0.2 to 0.4. This operation was described in detail in [11].

For the 2x1 setup, deteriorations of performance for all strategies are caused mainly by fading, while cross-correlation can be treated as a nearly small supplement. For 2x2 or 2x3 setups and ML strategy, the influence of both factors - fading and cross-correlation - are comparable, while for ZF and ML-ST the cross-correlation dominates.

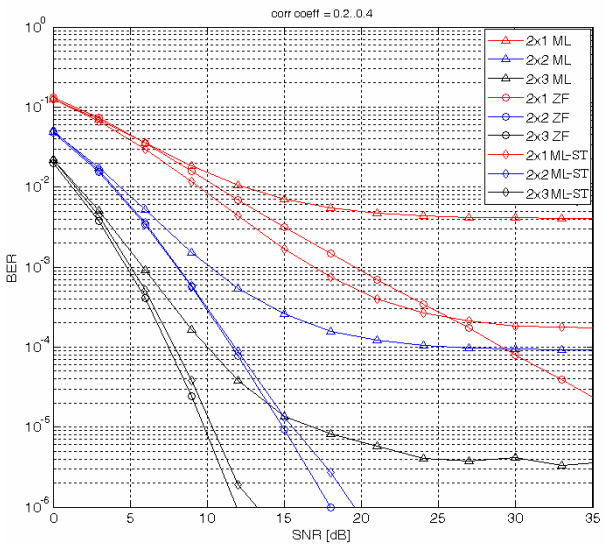


Figure 4. BER performance of ML, ZF, ML-ST strategies for setups 2x1, 2x2, 2x3 and fast fading, cross-correlation of ρ=0.2-0.4

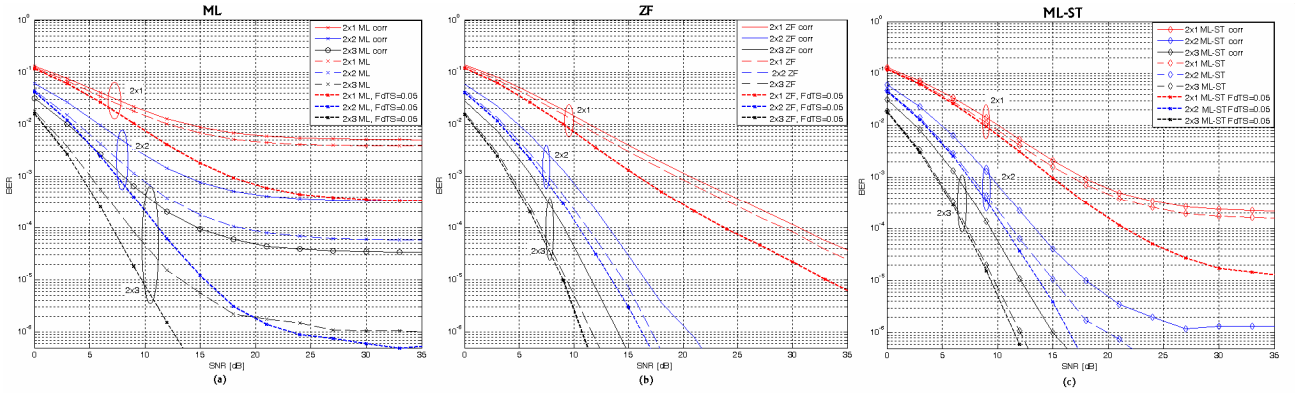


Figure 5. BER performance of ML (a), ZF (b) and ML-ST (c) strategies for strong correlated channels ρ= 0.5 - 0.7 and fast fading FdTS=0.1. Antennas setups 2x1, 2x2 and 2x3

However, it is worth noting that for the 2x3 setup and typical conditions of moderate fading and low cross-correlation, the simple ML detector offers better performance than ZF or ML-ST applied in a 2x1 or 2x2 setup, Fig. 3 and 4.

V. CONCLUSION

We carried out a comparison of ML, ZF, MMSE and ML-ST reception strategies in STBC MIMO systems for 2x1, 2x2 and 2x3 antennas setup. As a criterion of the comparison of the BER, characteristics obtained via Monte Carlo simulation were used. These characteristics show that an additional antenna (setup 2x3) used in a typical moderate fading and low cross-correlation environment gives better results than a sophisticated reception procedure, e.g. ML-ST. The benefits of using MMSE instead of ZF become smaller with an increase of the number of receiving antennas. The greatest impact on characteristics is caused by the number of receiving antennas. It was also shown that a low cross-correlation acts as an additive error source and is nearly independent on fading. The above conclusions are, however, not valid for a strong cross-correlation case and for fast fading where the ZF/ML-ST strategies bring even better results than ML in a 2x3 setup.

REFERENCES

[1] J. Foschini and M. J. Gan, “On Limits of Wireless Communications in a Fading Environment when using Multiple Antennas,” Wireless Personal Telecommunications, vol. 6, March 1998, pp. 311-335.

[2] I. E. Telatar, “Capacity of Multiantenna Gaussian Channels,” Bell Laboratories Technical Memorandum, June 1995.

[3] S. Alamouti, “A Simple Transmit Diversity Technique for Wireless Communications,” IEEE-JSAC, vol.16, No.8, 1998, pp 1451-1458.

[4] J. Jootar, J. R. Zeidler and J.G. Proakis, "Performance of Alamouti Space-Time Code in Time-Varying Channels with Noisy Channel Estimates', WCNC 2005, pp. 498-503.

[5] L. Y. Song and A. G. Burr, "Successive Interference Cancellation for Space-Time Block Codes over Time-Selective Channels," IEEE Comm. Letters, No. 12, December 2006, pp.837-839.

[6] F. C. Zheng and A. G. Burr, "Signal Detection for Orthogonal Space-Time Block Coding Over Time-Selective Fading Channels: The Hi Systems," IEEE Transactions on Communications, vol.5, No.1, January 2006, pp. 40-46.

[7] A. Vielmon, Y. Li, and J. R. Barry, "Performance of Alamouti Transmit Diversity Over Time-Varying Rayleigh Fading Channels," IEEE Transactions on Wireless Communications, vol.3, No.5, September 2004, pp. 1369-1373.

[8] J. Jootar, J. R. Zeidler, and J. G. Proakis, "Performance of Alamouti Space-Time Code in Time-Varying Channels with Noisy Channel Estimates,” WCNC 2005, pp. 498-503.

[9] P. C. Lozano, V. M. G. Yaez, and P. A. M., „Left-Pseudoinverse MIMO Channel Matrix Computation” Proceedings on International Conference on Electrical, Communications, and Computers 2009, pp. 134-138.

[10] L. Zhao and V. K Dubej, "Detection Schemes for Space-Time Block Code and Spatial Multiplexing Combined System," IEEE Comm. Letters, vol. 9, no. 1, January 2005, pp. 49-51.

[11] K. Kosmowski and J. Pawelec, “A Stochastic Model of Channel Correlation in MIMO Systems,” Military Communications and Information Systems Conf., Wrocław 2010, pp. 689-696.

[12] J.G. Proakis, Digital Communications, 4rd edition, *McGraw-Hill*, 2001 pp. 809;

[13] K.S. Alimgeer, A.Naveed, M. A. S. Chaudhry, and I. M. Qurashi, “A comparison of Interference Cancellation scheme based on QR decomposition and based on MMSE for Space Time Block code in multi-rate multi-user Systems,” Proceedings on IEEE Multitopic Conference, INMIC 2006, pp. 79-83.

 ISBN: 978-1-61208-126-7