ICN 2015

The Fourteenth International Conference on Networks

ISBN: 978-1-61208-398-8

SOFTNETWORKING 2015

The International Symposium on Advances in Software Defined Networks

April 19 - 24, 2015

Barcelona, Spain

ICN 2015 Editors

Carlos Becker Westphall, University of Santa Catarina, Brazil
Eugen Borcoci, University Politehnica of Bucharest, Romania
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland

ICN 2015

Foreword

The Fourteenth International Conference on Networks (ICN 2015), held between April 19th-24th, 2015 in Barcelona, Spain, continued a series of events focusing on the advances in the field of networks.

ICN 2015 welcomed technical papers presenting research and practical results, position papers addressing the pros and cons of specific proposals, such as those being discussed in the standard fora or in industry consortia, survey papers addressing the key problems and solutions, short papers on work in progress, and panel proposals.

ICN 2015 also featured the following Symposium:

- SOFTNETWORKING 2015: The International Symposium on Advances in Software Defined Networks

We take here the opportunity to warmly thank all the members of the ICN 2015 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICN 2015. 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 ICN 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICN 2015 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of networks.

We hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

ICN 2015 Advisory Committee:

Pascal Lorenz, University of Haute Alsace, France
Tibor Gyires, Illinois State University, USA
Carlos Becker Westphall, University of Santa Catarina, Brazil
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland

SOFTNETWORKING 2015 Advisory Committee:

Eugen Borcoci, University Politehnica of Bucharest, Romania
Pedro A. Aranda Gutiérrez, Telefónica, Spain
Nicola Ciulli, Nextworks, Italy
Marc Sune, Berlin Institute for Software Defined Networks GmbH, Germany
Wolfgang John, Ericsson Research, Sweden

ICN 2015

Committee

ICN Advisory Committee

Pascal Lorenz, University of Haute Alsace, France
Tibor Gyires, Illinois State University, USA
Carlos Becker Westphall, University of Santa Catarina, Brazil
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland

ICN 2015 Technical Program Committee

Alireza Abdollahpouri, University of Kurdistan - Sanandaj, Iran
Colin Allison, University of St Andrews, UK
Natalia Amelina, St. Petersburg State University, Russia | Norwegian University of Science and Technology, Norway
Jalel Ben-Othman, Université de Versailles, France
Max Agueh, LACSC - ECE Paris, France
Kari Aho, University of Jyväskylä, Finland
Pascal Anelli, Université de la Réunion, France
Cristian Anghel, Politehnica University of Bucharest, Romania
Jocelyn Aubert, Luxembourg Institute of Science and Technology (LIST), Luxembourg
Harald Baier, Hochschule Darmstadt, Germany
Alvaro Barradas, University of Algarve, Portugal
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Zdenek Becvar, Czech Technical University in Prague, Czech Republic
Djamel Benferhat, University of South Brittany, France
Ilham Benyahia, Université du Québec en Outaouais - Gatineau, Canada
Robert Bestak, Czech Technical University in Prague, Czech Republic
Jun Bi, Tsinghua University, China
Bruno Bogaz Zarpelão, State University of Londrina (UEL), Brazil
Patrick-Benjamin Bök, Weidmueller Group, Germany
Fernando Boronat Seguí, Universidad Politécnica de Valencia, Spain
Agnieszka Brachman, Silesian University of Technology - Gliwice, Poland
Arslan Broemme, GI BIOSIG - GI e.V., Germany
Matthias R. Brust, Technological Institute of Aeronautics, Brazil
Damian Bulira, Wroclaw University of Technology, Poland
Manoel Camillo Penna, Pontifícia Universidade Católica do Paraná, Brazil
Bin Cao, Harbin Institute of Technology Shenzhen Graduate School, China
Jorge Luis Castro e Silva, UECE - Universidade Estadual do Ceará, Brazil
Joaquim Celestino Júnior, Universidade Estadual do Ceará (UECE), Brazil
Eduardo Cerqueira, Federal University of Para, Brazil
Marc Cheboldaeff, T-Systems International GmbH, Germany
Buseung Cho, KREONET Center, KISTI - Daejeon, Republic of Korea

Konstantinos Patsakis, University of Piraeus, Greece
João Paulo Pereira, Polytechnic Institute of Bragança, Portugal
Kun Peng, Institute for Infocomm Research, Singapore
Ionut Pirnog, "Politehnica" University of Bucharest, Romania
Marcial Porto Fernandez, Universidade Estadual do Ceara (UECE), Brazil
Iwona Pozniak-Koszalka, Wroclaw University of Technology, Poland
Jani Puttonen, Magister Solutions Ltd., Finland
Shankar Raman, Indian Institute of Technology - Madras, India
Victor Ramos, UAM-Iztapalapa, Mexico
Priyanka Rawat, INRIA Lille - Nord Europe, France
Shukor Razak, Universiti Teknologi Malaysia (UTM), Malaysia
Yenumula B. Reddy, Grambling State University, USA
Eric Renault, Institut Mines-Télécom - Télécom SudParis, France
M. Elena Renda, IIT - CNR, Italy
Krisakorn Rerkrai, RWTH Aachen University, Germany
Karim Mohammed Rezaul, Glyndwr University - Wrexham, UK
Wouter Rogiest, Ghent University, Belgium
Simon Pietro Romano, University of Napoli Federico II, Italy
Angelos Rouskas, University of Piraeus, Greece
Jorge Sá Silva, University of Coimbra, Portugal
Teerapat Sanguankotchakorn, Asian Institute of Technology - Klong Luang, Thailand
Susana Sargento, University of Aveiro, Portugal
Panagiotis Sarigiannidis , University of Western Macedonia - Kozani, Greece
Masahiro Sasabe, Graduate School of Information Science - Nara Institute of Science and Technology, Japan
Thomas C. Schmidt, HAW Hamburg, Germany
Hans Scholten, University of Twente- Enschede, The Netherlands
Dimitrios Serpanos, ISI/RC Athena & University of Patras, Greece
Narasimha K. Shashidhar, Sam Houston State University - Huntsville, USA
Pengbo Si, Beijing University of Technology, P.R. China
Frank Siqueira, Federal University of Santa Catarina - Florianopolis, Brazil
Kamal Singh, Telecom Bretagne, France
Peter Skworcow, MontFort University - Leicester, UK
Karel Slavicek, Masaryk University Brno, Czech Republic
Andrew Snow, Ohio University, USA
Arun Somani, Iowa State University - Ames, USA
Kostas Stamos, University of Patras, Greece
Lars Strand, Nofas Management, Norway
Aaron Striegel, University of Notre Dame, USA
Miroslav Sveda, Brno University of Technology, Czech Republic
Maciej Szostak, Wroclaw University of Technology, Poland
Nabil Tabbane, SUPCOM, Tunisia
János Tapolcai, Budapest University of Technology and Economics, Hungary
Carlos Miguel Tavares Calafate, Universidad Politécnica de Valencia, Spain
Ken Turner, The University of Stirling, UK
Emmanouel Varvarigos, University of Patras, Greece
Dario Vieira, EFREI, France
Calin Vladeanu, University Politehnica of Bucharest, Romania

Matthias Vodel, Technische Universitaet Chemnitz, Germany
Lukas Vojtech, Czech Technical University in Prague, Czech Republic
Krzysztof Walkowiak, Wroclaw University of Technology, Poland
Boyang Wang, Xidian University, China
Tingka Wang, London Metropolitan University, UK
You-Chiun Wang, National Sun Yat-sen University, Taiwan
Yufeng Wang, University of South Florida - Tampa | NEC-Labs America - Princeton, USA
Gary Weckman, Ohio University, USA
Alexander Wijesinha, Towson University, USA
Maarten Wijnants, Hasselt University-Diepenbeek, Belgium
Bernd Wolfinger, University of Hamburg, Germany
Kok-Seng Wong, SoongSil University, South Korea
Qin Xin, University of the Faroe Islands, Faroe Islands
Lei Xiong, National University of Defense Technology - ChangSha, China
Qimin Yang, Harvey Mudd College-Claremont, USA
Vladimir Zaborovski, Polytechnic University of Saint Petersburg, Russia
Pavel Zahradnik , Czech Technical University Prague, Czech Republic
Morteza Mohammadi Zanjireh, Glasgow Caledonian University, UK
Arkady Zaslavsky, CSIRO ICT Centre & Australian National University - Acton, Australia
Sherali Zeadally, University of Kentucky, USA
Bing Zhang, National Institute of Information and Communications Technology - Yokosuka, Japan
Bo Zhao, Samsung Research America, USA
Tayeb Znati, University of Pittsburgh, USA
André Zúquete, IEETA - University of Aveiro, Portugal

SOFTNETWORKING 2015 Advisory Committee

Eugen Borcoci, University Politehnica of Bucharest, Romania
Pedro A. Aranda Gutiérrez, Telefónica, Spain
Nicola Ciulli, Nextworks, Italy
Marc Sune, Berlin Institute for Software Defined Networks GmbH, Germany
Wolfgang John, Ericsson Research, Sweden

SOFTNETWORKING 2015 Technical Program Committee

Pedro A. Aranda Gutiérrez, Telefónica, Spain
Robert Bestak, Czech Technical University in Prague, Czech Republic
Eugen Borcoci, University Politehnica of Bucharest, Romania
Nicola Ciulli, Nextworks, Italy
Didier Colle, iMinds - Ghent University, Belgium
Daniel Corujo, University of Aveiro, Portugal
Christian Esteve Rothenberg, University of Campinas, Brazil
Roberto Gonzalez, NEC Laboratories Europe, Germany
Wolfgang John, Ericsson Research, Sweden
Eiji Kawai, National Institute of Information and Communications Technology, Japan
Wolfgang Kiess, DOCOMO Euro-Labs, Germany

Zoltán Lajos Kis, Ericsson, Hungary
Diego Kreuz, University of Lisboa , Portugal
Maciej Kuzniar, EPFL, Switzerland
Luca Prete, Open Networking Laboratory (ON.LAB), USA
Milosz Przywecki, Poznan Supercomputing and Networking Center, Poland
Fernando Ramos, University of Lisbon, Portugal
Yang Song, IBM Almaden Research Center, USA
Marc Sune, Berlin Institute for Software Defined Networks GmbH, Germany
Yutaka Takahashi, Kyoto University, Japan
Samir Tohmé, PRISM Laboratory - University of Versailles, France
Jozef Wozniak, Gdańsk University of Technology, Poland

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Priority Levels in a HIPERLAN Based Forwarding Mechanism for Intermittent Connectivity

Constantine Coutras
Department of Computer Science
Montclair State University
Montclair, NJ 07043
Email: coutrasc@montclair.edu

***Abstract*—As the proliferation of wireless devices continues to outpace the necessary infrastructure to support them everywhere, intermittent connectivity is becoming common in certain wireless access networks. Today's existing network protocols are not resilient to disruption of communication links and communication often fails when faced with sporadic connectivity. Thus, there is a growing interest in Intermittently Connected Mobile Networks (also know as Mobile Opportunistic Networks) and a number of routing protocols and frameworks have been proposed in the literature to address the problem. Different priority classes, each containing different priority levels are introduced. The paper concludes with numerical results from simulation, which takes into account bit errors, hidden nodes and capture.**

***Keywords* - Wireless Networks; Intermitent Connectivity; Forwarding Mechanism.**

I. INTRODUCTION

The ongoing evolution of wireless systems has created a world demanding wireless connectivity everywhere. Most wireless access today is achieved through wireless LANs, operating mostly in infrastructure mode and to a lesser degree in Ad Hoc, smartphones and tablets. But there are still situations where such connectivity is not possible. But wireless access networks worldwide fail to fulfill the promise of continuous, high-bandwidth, and affordable service everywhere.

Intermittently connected mobile networks (ICMNs) are networks where most of the time, there does not exist a complete end-to-end path from the source to the destination. Even if such a path exists, it may be highly unstable because of topology changes due to mobility. Thus, traditional wireless technologies that require an end to end path are not suitable for these networks. There is a growing interest in these networks and many routing protocols have been proposed in the literature to address the problem. Publications presenting aspects of these networks include [1][2][3]. Extremely Opportunistic Routing (ExOR), a unicast routing technique for multihop wireless networks is presented in [4][5]. A middleware design called the Self Limiting Epidemic Forwarding (SLEF) [6], automatically adapts its behavior from single hop MAC layer broadcast to epidemic forwarding when the environment changes from being extremely dense to sparse, sporadically connected. Prioritized Epidemic Routing (PREP) is described in [7]. And lately, FansyRoute is introduced in [8], and a probability-based "Spray and Wait protocol" in [9].

A thorough analysis of different techniques used for routing (direct routing, epidemic routing, randomized flooding and spraying techniques) that also includes contention for the wireless channel in the analysis can be found in [10][11][12]. Finally, the problem of accurately modeling mobility and using realistic traces of mobility in these networks is analyzed in [13]–[16].

The High PErformance Radio Local Area Network (HIPERLAN) has been developed by the European Telecommunications Standards Institute (ETSI). The HIPERLAN protocol functionality is presented in [17]. Other sources of information and analysis of the HIPERLAN protocol can be found in [18]–[24]. A study of the protocol's performance for asynchronous traffic that takes into account the phenomena of hidden nodes and capture is presented in [25] and a study of the protocol's performance for real-time traffic that takes into account the phenomena of hidden nodes and capture is presented[26].

As an alternative to the approaches mentioned above for dealing with intermittent connectivity, we wish to provide a solution for intermittent connectivity based on modifying an existing wireless protocol (HIPERLAN) so that it operates in the same exact way either the wireless environment in one of intermittent connectivity or not. This modification stays at the medium access layer and thus does not involve a routing algorithm which would be found at a higher layer. Thus, this approach differs from the ones referenced above and a one to one comparison is not saught.

In this paper we extend our previous work, found in [27] on modifying the HIPERLAN Channel Access Control (CAC) Layer protocol to accommodate for loss of connectivity, by introducing two classes of service, each containing different priority levels. In Section 2, an overview of the HIPERLAN CAC Layer Protocol is given in which channel access in the presence of hidden nodes is also taken into account. In Section 3, our modification of the HIPERLAN CAC Layer for intermittently connected mobile networks is presented. We then discuss packet transmission in the presence of hidden nodes, capture and bit errors. In Section 4, classes of service and priority levels are introduced. In Section 5, numerical results from simulation study are presented. And in Section 6, the conclusion is given with directions for future work.

 ISBN: 978-1-61208-398-8

II. Overview of the HIPERLAN CAC Layer Protocol

The CAC layer is actually the "lower sublayer" of the MAC layer that basically deals with the mechanism of accessing the channel (EY-NPMA mechanism).

The three phases of the EY-NPMA mechanism constitute the contention phase of the *Synchronized Channel Access Cycle*.

In Figure 1 we see a renewal interval, its components, and their components as well. Transmission is denoted by black color, while its absence is denoted by white color, and a different shade filling is used for the synchronization slot.

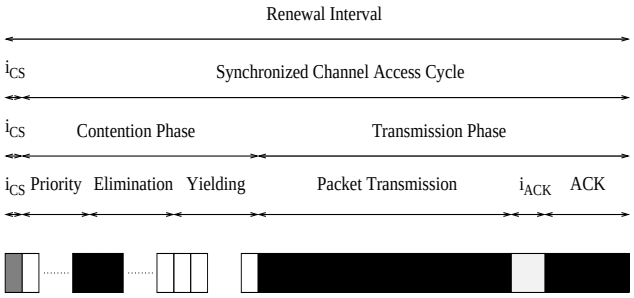


Fig. 1. The EY-NPMA mechanism.

In the prioritization phase, an active node of priority n must signal its intention to access the channel by transmitting a burst during the nth time slot, provided that no active nodes of higher priority have already signaled their intention to access the channel. A total of m_{CP} priority levels, from 0 to $m_{\mathrm{CP}}-1$ is assumed.

In the elimination phase, each active node bursts a signal for a random number of time slots and then listens to the channel; if another active node is still bursting this active node is eliminated, otherwise it may continue into the yielding phase. In the elimination phase we have a maximum of m_{ES} elimination slots. The probability of bursting in an elimination slot is p_{E} . The maximum burst allowed is $m_{\mathrm{ES}}-1$ time slots.

In the yielding phase, each active node listens for a random number of time slots and then, if the channel is still free, starts a packet transmission. In the yielding phase we have a maximum of m_{YS} yielding slots. The probability of yielding in a yielding slot is p_{Y} . An active node can listen for a maximum of $m_{\mathrm{YS}}-1$ time slots.

A. Channel Access in the Presence of Hidden Nodes

Due to the high channel speed used in HIPERLAN we can assume relatively limited *node mobility*. This means that if two nodes are hidden from each other in the beginning of a renewal interval they will remain hidden through out that renewal interval.

During the prioritization phase, for an active node of lower priority to falsely determine that it may continue into the next phase, it will have to be hidden from all active nodes of higher priorities.

From [25][26] we have that an active node might falsely determine that it has survived the elimination phase if it is hidden from all active nodes that burst for more time slots that it does; while in the yielding phase, a node might falsely determine channel access.

After the elimination phase is over, it is possible that not all nodes, which a node i "sees" winning the elimination phase, will continue into the next phase. This is due to the fact that some of them could have been eliminated by other -non hidden from them- nodes, that are hidden from node i. It is also possible that during the yielding phase, even more nodes are eliminated before i wins channel access due to the fact that they hear transmission from other nodes not hidden from them but -again- hidden from i. A detailed calculation of the probability of channel access for a node can be found in [25].

III. Modifying the HIPERLAN CAC Layer for Intermittently Connected Mobile Networks

Since a complete and stable end-to-end path between source and destination does not exist in intermittently connected mobile networks (ICMNs), packets need to be stored and then forwarded on an evolving path from source to destination. As shown in Figure 1, after the EY-NPMA mechanism determines the node(s) that can use the channel to transmit a packet, the packet and acknowledgment are expected to be directly exchanged between source and destination. In order to allow for the packet to be forwarded by other nodes to its destination, these other nodes should be allowed to acknowledge receipt of the packet after they have determined that the destination has not successfully received it while they have. These nodes will then take it upon themselves to forward the packet towards its destination. A key decision that has to be made is which nodes with direct access to the transmitter should forward the packet. Allowing too many of them would increase the offered load on the channel and decrease performance. On the other hand if not enough nodes take on the task of forwarding the packet, its total time to the destination is going to increase on average and the chances of the packet following an optimal path will also decrease. We propose that only the nodes that were actively competing for channel access and lost it should listen for the packet and acknowledge it if they have successfully received it. They should then add it to the list of packets that they need to forward. This would limit the number of forwarding nodes to the ones already competing, thus not introducing any new nodes into channel competition. When one of these nodes will eventually win channel access, it will transmit its own packet and then the packets that it has taken upon to forward. In addition, if a node receives for forwarding a packet that it has already forwarded, it will not forward it, but discard it as a duplicate. Each packet should also have a "time to live" TTL, so that packets that are taking a long path towards the destination are eliminated (some packets might never reach destination).

 ISBN: 978-1-61208-398-8

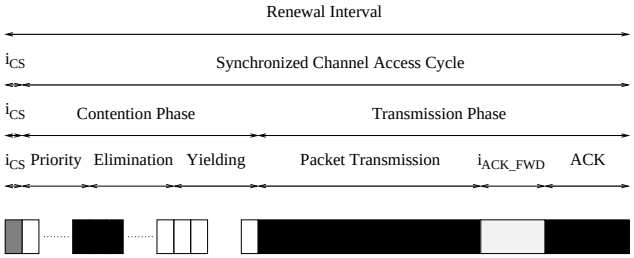


Fig. 2. Allowing acknowledgments from forwarding nodes.

Another key decision that has to be made, is the buffer size that each node will allocate for forwarding packets. With a larger buffer space more packets can be buffered for forwarding, but at the risk of increasing the number of extra copies of each packet that needs to be forwarded (increase in total traffic).

The proposed modifications which are necessary to implement this forwarding mechanism are the following:

- All nodes that unsuccessfully contended for channel access should listen for the transmission of a packet by the winning node. If they determine it to be successful (by hearing an acknowledgment after a time interval of i_{ACK}) they need to do nothing else. But If they hear no acknowledgment after a time interval of i_{ACK} they should acknowledge the packet at a time interval of $i_{\mathrm{ACK\_FWD}}$ from its transmission, where $i_{\mathrm{ACK\_FWD}}$ should obviously be longer than i_{ACK}. Figure 2 shows such an acknowledgment sent by a forwarding node.
- If no acknowledgment is received after the time interval of $i_{\mathrm{ACK\_FWD}}$, the transmission has failed and the renewal interval is completed.
- If the transmitter node receives acknowledgments in a time interval of $i_{\mathrm{ACK\_FWD}}$ it determines that its transmission is successful and its packet will be forwarded. This should be true even if multiple overlapping acknowledgments are received.
- A node that successfully transmits a packet of its own can continue in the same renewal interval to transmit packets it has taken upon to forward, up to a maximum number of "forwarding transmissions" of FT_{MAX}. This modified transmission phase is shown in Figure 3, where $i_{\mathrm{ACK}*}$ could either be i_{ACK} or $i_{\mathrm{ACK\_FWD}}$ and i_{NT} (which is less than i_{CS}) is the time interval between transmissions.
- As soon as all forwarding transmissions are done or one of the extra forwarding transmissions fails the renewal interval is over.

We also propose that the total priority levels are only two: high and low. All nodes start out as low priority nodes. It is expected that traffic that will be allowed to switch to the higher priority will have better quality of service. Alternative methods for elevating a nodes priority are presented in the next section.

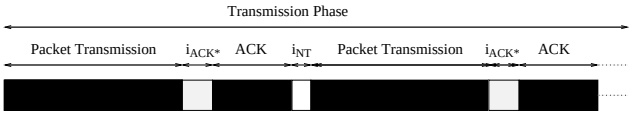


Fig. 3. Modified Transmission Phase.

A. Packet Transmission in the Presence of Hidden Nodes, Capture and Bit Errors

We start our analysis of packet transmission in the presence of hidden nodes, capture and bit errors by examining the unmodified HIPERLAN CAC Layer operation.

Let d_{ij} denote the distance between two nodes i and j. A transmission by station i will be "captured" by station j if no other node in a circle of radius αd_{ij}, $\alpha \geq 1$, around j is transmitting simultaneously. The parameter α will be referred to as the *capture parameter* and the area surrounding node j of radius αd_{ij} will be referred to as $j's$ *capture area* for node i. Let's further denote $A_{ji}(\alpha)$ as the set of nodes that are in $j's$ capture area for i, and $\bar{A}_{ji}(\alpha)$ the set of nodes that are out of $j's$ capture area for i.

The assumption of relatively limited node mobility is extended, so that if a node is inside or outside of a specific capture area it will remain so throughout the renewal interval.

In our modified HIPERLAN CAC Layer, we allow in between nodes to receive a packet, and if the destination node has not successfully received it, to store it for forwarding, and to acknowledge it. Thus, the conditions for a node to successfully transmit a packet and successfully receive its acknowledgment are now the following:

- The transmitter and destination nodes are not hidden from each other or the transmitter and a potential forwarding node are not hidden from each other.
- After the contention phase is over, the transmitter node determines channel access, while the destination node, if active and not hidden from the transmitter node and with a direct link to it, does not.
- If nodes other than the transmitter node have survived the contention phase, they have to be either hidden from the destination node or a potential forwarding node or otherwise they have to be outside of the destination's capture area for the transmitter or of a potential forwarding node's capture area for the transmitter, so that the transmitter's packet can be successfully received.
- No node that is not hidden from the transmitter or that is not outside of the transmitter's capture area for the destination, receives successfully a packet if a direct link between transmitter and destination exists.
- If a direct link between transmitter and destination node does not exist no node that is not hidden from the transmitter or that is not outside of the transmitter's capture area for a receiving forwarding node, receives successfully a packet.
- The destination or a forwarding node receives a bit error free packet.

 ISBN: 978-1-61208-398-8

IV. Classes of Service and Priority Levels

In ICMNs it is expected that nodes that will not experience the same quality of service that they experience in ordinary fully connected WLANs. Thus, giving priority to some traffic over other will increase the quality of service it experiences.

Different ways of controlling priority elevation can be used. If no classes of service exist and all traffic from all nodes is treated equal, a simple mechanism can be in place to elevate the priority of packets that have been experiencing delays trying to get through to their destination. If a node has failed to gain channel access for a specific number of consecutive attempts (denoted by LS_{MAX}) and it has packets in addition to its own to forward, then its priority increases to the high priority. Once channel access is granted it returns to the low priority after a certain number of packets have been transmitted.

We further propose to allow for different classes of service. Nodes that have subscribed to a higher class of service have access to higher priority levels. Each class of service should further have a "high" and a "low" priority level. Packets can then transition from the lower to the higher priority for their class based on the mechanism described above. This allows for a total of four priorities, where the top two are only accessed by higher class of service nodes and the bottom two are only accessed by the lower class of service.

V. Simulation Study and Numerical Results

The simulation of the proposed forwarding algorithm takes into account realistic conditions for wireless intermittent connectivity (contention, mobility, cluster formations, bit errors, hidden nodes and capture).

To evaluate the proposed forwarding algorithm we start with 25 wireless nodes that need to communicate with each other as in an ad-hoc wireless network, but each node can have direct transmission only to nodes in a limited area surrounding it, thus the nodes in that area (or cluster) that successfully receive a transmission from that node will have to forward it to other nodes outside this area. Then mobility is added that will allow for these clusters to change over time and communication to be realized over a dynamically evolving path. The mobility model for each node is that of a *random walk*. The maximum velocity allowed is 100km/h, to account for a wireless device in a vehicle. The simulation is run for various cluster area sizes. We also assume that the probability that two nodes will be hidden from each other for reasons other that being out of range from each other (in different clusters) to be a typical $p_h = 0.02$ and we also take into account capture but with a capture parameter of $\alpha = 10$, which introduces a very limited capture benefit. The bit error rate is $BER = 10^{-6}$, which is typical for wireless transmissions. The initial coordinates of the nodes were chosen randomly and are normalized to belonging inside a circle of radius 1. These nodes are then allowed to move totally randomly in a square defined by the points (-1,1), (1,1), (1,-1) and (-1,-1). The initial coordinates at the beginning of the simulation are chosen randomly.

It is assumed for simplicity that all sources generate packets following the Poisson distribution (inter-arrival times between packets follow the exponential distribution). And all nodes are transmitting the same amount of load and equally to all other nodes. It is also assumed that all nodes initially are of the same low priority level for the class of service they belong to. The new protocol parameter used is: $i_{\mathrm{ACK\_FWD}} = 384$ bits (while $i_{\mathrm{ACK}} = 256$ bits). The channel bandwidth for HIPERLAN/1 is 23529 Kbits/sec.

Various values were chosen for FT_{MAX}, LS_{MAX}, and the maximum buffer size allocated for receiving packets to forward FB_{MAX}. AS with previous work, we will present results for $FT_{\mathrm{MAX}} = 8$, and $FB_{\mathrm{MAX}} = 20$.

Finally, each packet has a time to live of $TTL = 10$. This value is smaller than the typical value found in other protocols, but given that a great number of nodes will be forwarding a packet as load increases in the network, a smaller TTL value will help in keeping network load from increasing to a point of severe congestion.

A. Numerical Results

We are interested in the expected delay between initial transmission by the source and the final acknowledgement by the destination. For various values of load and for various values of the cluster area we plot the probability of overflow. The probability of overflow is the probability of a packet exceeding a maximum time delay value, and is the main focus of this simulation study.

In our previous work it was shown that the expected delays with no classes of service were acceptable and in some cases good enough for real time traffic. In this simulation study we have two different classes of service and each class of service has two different priority levels so that in each class of service fairness can be achieved and similar performance achieved.

The data for the lower class of service show that this class of service experiences increased delays in order for the higher class of service to achieve better performance. Thus, the lower class of service is not suitable for real time traffic. In this section we will concentrate in presenting data for the high priority class. All the data presented is for a simulation of 5 nodes being of higher class of service while the other 20 remain at the lower class of service. Other configurations were also explored, but as the number of nodes of higher priority increased the benefit to the higher priority deminished. Thus, the number of nodes of higher priority must be limited.

In Figure 4 one can see the plot of the probability of overflow for various loads. These loads are between 10% and 20% of the load capacity of the channel. All loads are for a cluster area = 9.61% Better performance is achieved at higher loads as more nodes are forwarding packets at higher loads. This is something that we also saw with the simulation of only one class of service.

 ISBN: 978-1-61208-398-8

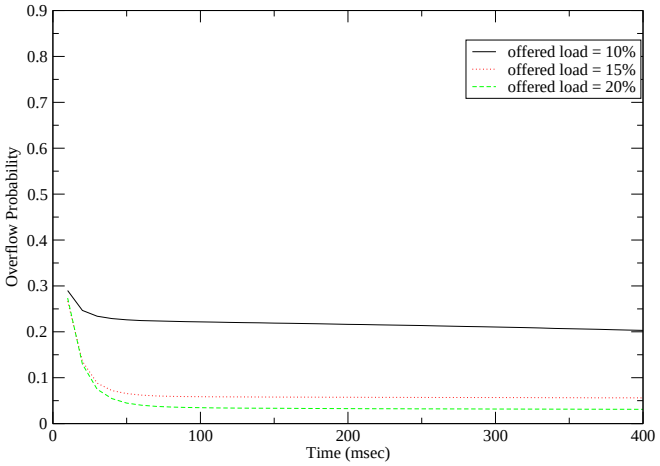


Fig. 4. Probability of overflow for cluster area = 9.61%

In Figure 5 one can see the plot of the probability of overflow for various loads, again between 10% and 20% of the load capacity of the channel. The cluster area has increased now to 12.56% and the performance is about the same. One interesting difference is that with the increase of cluster area we see that performnce is worse for a load of 20% compared to a load of 15%. Thus, the increase in load while initially increased performance, after a certain load increase, started to decrease performance in this cluster size.

In Figure 6 one can see another plot of the probability of overflow for various loads, again between 10% and 20% of the load capacity of the channel. The cluster area has increased now to 15.90% and the performance has slightly dropped as delays have slightly gone up. This can be attributed to the fact that with a bigger cluster area more competition exists to forward a packet to the next cluster area.

And in Figure 7 one can see another plot of the probability of overflow for the various loads, again between 10% and 20% of the load capacity of the channel. The cluster area has now

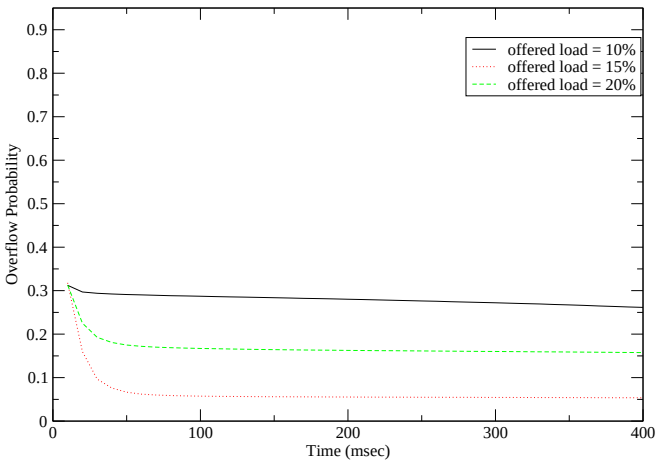


Fig. 5. Probability of overflow for cluster area = 12.56%

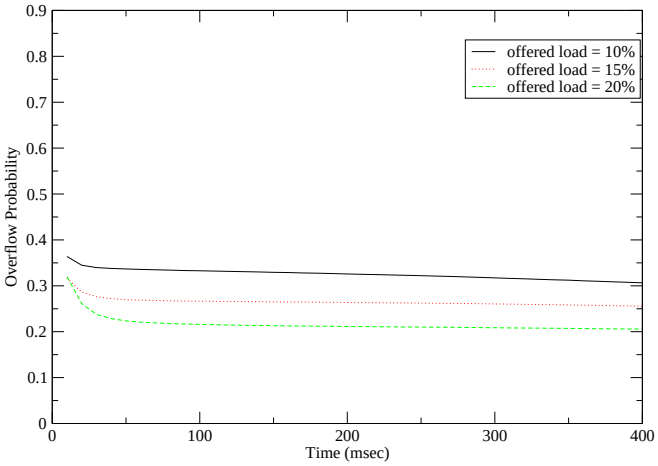


Fig. 6. Probability of overflow for cluster area = 15.90%

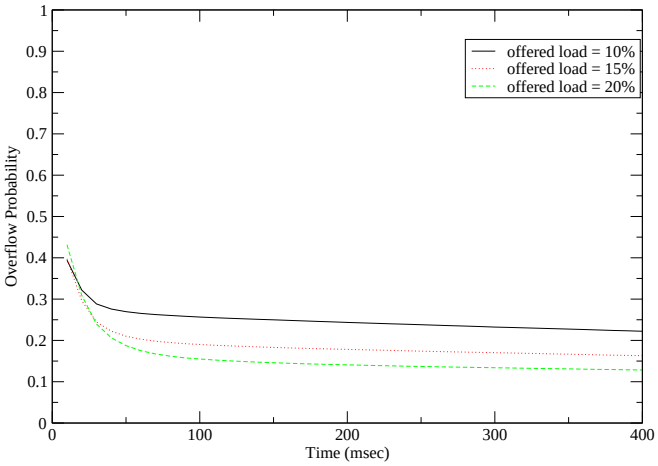


Fig. 7. Probability of overflow for cluster area = 23.74%

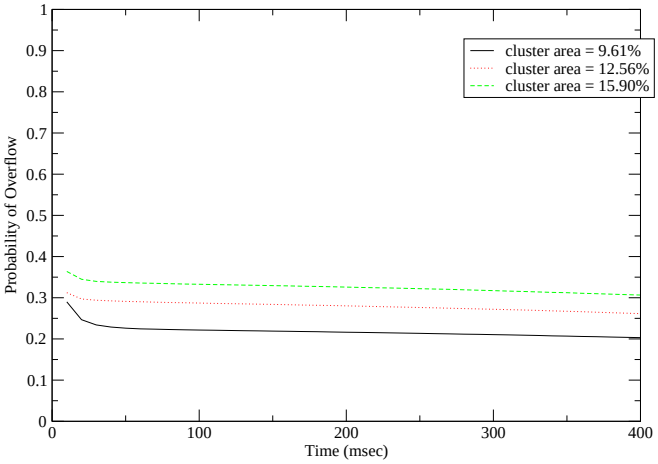


Fig. 8. Probability of overflow for load = 10%

 ISBN: 978-1-61208-398-8

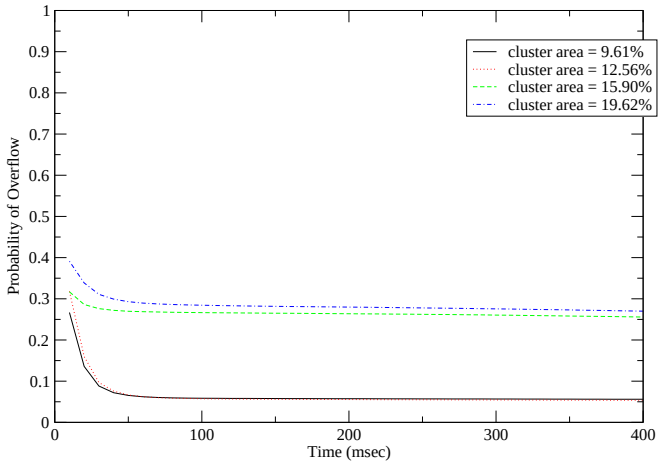


Fig. 9. Probability of overflow for load = 15%

increased to 23.74% and the performance has not decreased. We also see that performance differences between loads have come smaller.

In Figure 8 one can see the plot of the probability of overflow for various cluster areas. The cluster areas are between 9.61% and 15.90%. The load is at 10% of the channel capacity. In this figure we see similar performance for all three cluster areas, with better performance at the larger cluster area.

In Figure 9 the load has increased to 15% of the channel capacity and we again look at the plot of the probability of overflow for various cluster areas. We see that the higher load has increased performance at the lower cluster areas. Similar observation was made in our previous studies with no classes of service.

In Figure 10 one can see again the plot of the probability of overflow for various cluster areas. This time he load has increased to 20% and performance remains strong.

If we compare the above results with the results of the analysis of of different techniques used for routing (direct routing, epidemic routing, randomized flooding and spraying techniques) found in [10][11][12], we see that we obtain similar or almost as good results, although in those studies focus mainly on the expected delay rather than on the probabilty of exceeding a maximum delay. And once again this is done by only modifying the wireless access mechanism rather than introducing a routing algorithm. Thus, operating without the computational overhead a routing algorithm would introduce.

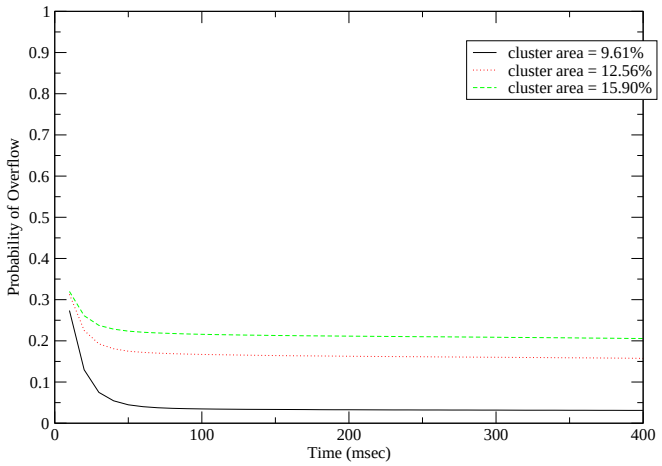


Fig. 10. Probability of overflow for load = 20%

VI. Conclusion

A brief outline of the HIPERLAN CAC layer protocol is presented. Then a modification to its operation that includes different classes of service is proposed and that would allow it to operate in ICMNs. The influence of the phenomena of hidden nodes, capture and bit errors is discussed and the conditions for successful packet transmission are presented for the modified forwarding mechanism. From the numerical data presented it is shown that the higher class of service can enjoy good enough performance to allow for real time loss tolerant applications. As the data for the lower class of service showed performance that would allow only non-real time traffic with substantial delays in some cases, it is not presented. This research is now progressing into deeper understanding of the performance data collected through simulation. One direction aims at further understanding the relationship between cluster areas and loads so a connection admission control mechanism can be created to ensure some minimum quality of service for the nodes of higher class service. A second direction is looking into how loss tolerant applications can further communicate their minimum needs of performance and affect the overall network performance.

References

[1] K. S. Phanse and J. Nykvist, “Opportunistic wireless access networks,” in AcessNets ’06: Proceedings of the 1st international conference on Access networks. New York, NY, USA: ACM, 2006, p. 11.

[2] D. Zheng, W. Ge, and J. Zhang, “Distributed opportunistic scheduling for ad-hoc communications: an optimal stopping approach,” in MobiHoc ’07: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing. New York, NY, USA: ACM, 2007, pp. 1–10.

[3] J. Kim and S. Bohacek, “A comparison of opportunistic and deterministic forwarding in mobile multihop wireless networks,” in MobiOpp ’07: Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking. New York, NY, USA: ACM, 2007, pp. 9–16.

[4] S. Biswas and R. Morris, “Opportunistic routing in multi-hop wireless networks,” SIGCOMM Comput. Commun. Rev., vol. 34, no. 1, 2004, pp. 69–74.

[5] ——, “Exor: opportunistic multi-hop routing for wireless networks,” in SIGCOMM ’05: Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications. New York, NY, USA: ACM, 2005, pp. 133–144.

[6] A. El Fawal, J.-Y. Le Boudec, and K. Salamatian, “Multi-hop broadcast from theory to reality: practical design for ad hoc networks,” in Autonomics ’07: Proceedings of the 1st international conference on Autonomic computing and communication systems. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007, pp. 1–10.

[7] R. Ramanathan, R. Hansen, P. Basu, R. Rosales-Hain, and R. Krishnan, “Prioritized epidemic routing for opportunistic networks,” in MobiOpp ’07: Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking. New York, NY, USA: ACM, 2007, pp. 62–66.

 ISBN: 978-1-61208-398-8

[8] S. Dabideen and R. Ramanathan, "Fansyroute: Adaptive fan-out for variably intermittent challenged networks," SIGMOBILE Mob. Comput. Commun. Rev., vol. 18, no. 1, Feb. 2014, pp. 37–45. [Online]. Available: http://doi.acm.org/10.1145/2581555.2581561

[9] E.-H. Kim, J.-C. Nam, J.-I. Choi, and Y.-Z. Cho, "Probability-based spray and wait protocol in delay tolerant networks," in 2014 International Conference on Information Networking (ICOIN), Feb 2014.

[10] A. Jindal and K. Psounis, "Contention-aware analysis of routing schemes for mobile opportunistic networks," in MobiOpp '07: Proceedings of the 1st international MobiSys workshop on Mobile opportunistic networking. New York, NY, USA: ACM, 2007, pp. 1–8.

[11] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Efficient routing in intermittently connected mobile networks: the multiple-copy case," IEEE/ACM Trans. Netw., vol. 16, no. 1, 2008, pp. 77–90.

[12] A. Jindal and K. Psounis, "Contention-aware performance analysis of mobility-assisted routing," Mobile Computing, IEEE Transactions on, vol. 8, no. 2, feb. 2009, pp. 145 –161.

[13] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Performance analysis of mobility-assisted routing," in MobiHoc '06: Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing. New York, NY, USA: ACM, 2006, pp. 49–60.

[14] L. Song and D. F. Kotz, "Evaluating opportunistic routing protocols with large realistic contact traces," in CHANTS '07: Proceedings of the second ACM workshop on Challenged networks. New York, NY, USA: ACM, 2007, pp. 35–42.

[15] M. Abdulla and R. Simon, "Characteristics of common mobility models for opportunistic networks," in PM2HW2N '07: Proceedings of the 2nd ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks. New York, NY, USA: ACM, 2007, pp. 105–109.

[16] V. Erramilli and M. Crovella, "Forwarding in opportunistic networks with resource constraints," in CHANTS '08: Proceedings of the third ACM workshop on Challenged networks. New York, NY, USA: ACM, 2008, pp. 41–48.

[17] HIgh PErformance Radio Local Area Network Type 1: functional specification, pr ets 300 652 ed., European Telecommunication Standards Institute, 650 Route des Lucioles, Sophia Antipolis, Valbonne, France, May 1996.

[18] G. Anastasi, L. Lenzini, and E. Mingozzi, "Stability and performance analysis of hiperlan," Proceedings of the IEEE Infocom, 1998.

[19] S. Chevrel, A. Aghvami, H. Lach, and L. Taylor, "Analysis and optimasation of the hiperlan channel access contention scheme," Wireless Personal Communications, vol. 4, no. May, 1996, pp. 27–39.

[20] K. Fu, Y. Guo, and S. Barton, "Performance of the ey-npma protocol," Wireless Personal Communications, vol. 4, no. May, 1996, pp. 41–50.

[21] P. Jacquet, P. Minet, P. Muhlethaler, and N. Rivierre, "Data transfer for hiperlan," Wireless Personal Communications, vol. 4, no. May, 1996, pp. 65–80.

[22] ——, "Priority and collision detection with active signaling - the channel access mechanism of hiperlan," Wireless Personal Communications, vol. 4, no. May, 1996, pp. 11–25.

[23] W. M. Moh, D. Yao, and K. Makki, "Wireless lan: Study of hidden-terminal effect and multimedia support," Proceedings of the IEEE Infocom, 1998, pp. 422–431.

[24] R. O. LaMaire, A. Krishna, P. Bhagwat, and J. Panian, "Wireless lans and mobile networking: Standards and future directions," IEEE Communications Magazine, no. August, 1996, pp. 86–94.

[25] C. Coutras and P.-J. Wan, "Evaluating performance of the HIPERLAN CAC layer protocol for asynchronous traffic," in Proceedings of the 24th Annual IEEE Conference on Local Computer Networks, 1999.

[26] C. Coutras, O. Frieder, and P.-J. Wan, "Evaluating performance of the HIPERLAN CAC layer protocol for real-time traffic," in Proceedings of the 25th Annual IEEE Conference on Local Computer Networks, 2000.

[27] C. Coutras, "Modifying the hiperlan/1 cac layer protocol for intermittent connectivity," in Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems, as part of the 2nd International Workshop on Mobile Computing and Networking Technologies, 2010.

 ISBN: 978-1-61208-398-8

Fine Angle Estimation Using Weighted Average-ESPRIT for Radar-based WSN

Sangdong Kim, Yeonghwan Ju, Daegun Oh, Jonghun Lee
Daegu Gyeongbuk Institute of Science & Technology
Advanced Radar Technology (ART) Lab., Robotics System Research Division
Daegu, the republic of Korea
e-mail: kimsd728@dgist.ac.kr, yhju@dgist.ac.kr, dgoh@dgist.ac.kr, jhlee@dgist.ac.kr

***Abstract*—This paper proposes a fine angle estimation using weighted average-estimation of signal parameters via rotational invariance techniques (ESPRIT) for radar-based WSN. The proposed WA-ESPRIT system is composed of a weighted average block and an ESPRIT block. The proposed system is verified through analysis, simulation and experiment. We show that the proposed system has better performance than the conventional one. The performance of the proposed algorithm is verified through Monte-Carlo simulations in an additive white Gaussian noise (AWGN).**

Keyword-radar-based WSN; Monitoring system; Weighted-average scheme; ESPRIT.

I. INTRODUCTION

Historically, surveillance systems have used infrared, acoustics and magnetics for passive sensing, and optics and ultrasounds for active sensing, but radio detection and ranging (radar) has been conspicuously absent. Conventional radar systems such as the pulse Doppler radar and the frequency modulated continuous-wave (FMCW) radar employ transmitted and reflected microwaves to detect, locate, and track objects over long distances and large areas. Due to its ability, radar has found applications in defense and remote sensing. However, widespread commercial applications of the radar have been limited because conventional systems are expensive, bulky and difficult to use. The radar motion sensors which have a short range and poor false alarm rates is used in unstructured environments such as traffic monitoring and police radar [1]. Since many wireless sensor networks (WSN) operate in unstructured environments with limited energy supplies, these conventional sensors are unsuitable for WSN. The FMCW radar is mostly used to gather traffic data information such as lane position and for radar-based WSN because the radar detects the angle between the target and the radar. The advantage of the FMCW radar is to obtain simple hardware architecture, lower peak level and low-cost comparing with pulse radar [2].

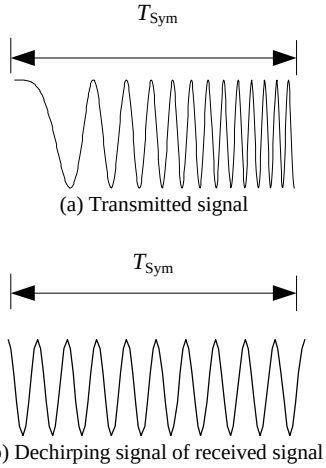
Among various estimated parameter such as distance and angle in radar detectors, the angle estimation is focused because the angle error is directly depended on the lane detection of the vehicle. Conventional angle estimation method show that parametric methods such as MUSIC, estimation of signal parameters via rotational invariance techniques (ESPRIT) and matrix pencil (MP) are used for super-resolution frequency estimation algorithms [3-5]. However, since conventional super-resolution method has assumed a lot of antenna arrays, the conventional one cannot operate well in case of a few arrays such as 2 or 3 arrays. Therefore, we propose weighted average (WA)-ESPRIT for low complexity realization of high resolution radar-based WSN. This paper is organized as follows. In Section 2, we introduce the system model for FMCW radar. Section 3 proposes WA-ESPRIT for the angle estimation, while Section 4 discusses the performance analysis of the WA-ESPRIT radar based on various parameters. Section 5 shows the simulation results for WA-ESPRIT radar and Section 6 shows experiments to testify the effectiveness of the proposed estimation. Finally, Section 7 concludes our proposed estimation with high resolution under real channel.

II. SYSTEM MODEL

The signal model of FMCW radar is expressed in this section. The transmitted FMCW chirp signal can be represented by

$$s(t)=\begin{cases}\exp\left[j\left((\omega_s+\omega_c)t+\frac{\mu}{2}t^2\right)\right] & \text{for } 0\le t<T_{sym}\\ 0 & \text{elsewhere}\end{cases}. \quad \textbf{(1)}$$

where ω_s denotes the start frequency, ω_c denotes the carrier frequency, μ is the rate of change of the instantaneous frequency of a chirp signal, and T_{sym} is the duration of chirp signals. The relation between the bandwidth of FMCW transmitted signal and μ is expressed by $\omega_{BW}=\mu T_{sym}$.



(a) Transmitted signal

(b) Dechirping signal of received signal

Figure 1. The signal scheme of FMCW radar signal

Consider M targets receiving at K antenna arrays. Let ϕ_m and τ_m denote the angle and delay of target of the m-

th target. The received signal at each antenna array can be represented by

$$r_k(t) = \sum_{m=0}^{M-1} a_m \exp\left(j\frac{2\pi}{\lambda} d(k-1)\sin\phi_m \right) s(t-\tau_m) + w_k(t) \quad (2)$$

where a_m denotes the complex amplitude for the m-th target, λ denotes the wave-length of the carrier signal, d is the spacing between the adjacent antenna elements, and $w_k(t)$ is the additive white Gaussian noise (AWGN) signal at the k-th antenna element.

In FMCW radar, received chirp signals can be easily transformed into the sinusoidal waveform by de-chirping as shown in Figure 1. Omitting AWGN signal, the sinusoidal signals of the received signal can be represented by

$$y_k(t) = \sum_{m=0}^{M-1} a_m \exp\left(j\frac{2\pi}{\lambda} dk\sin\phi_m \right) \exp\left(j\left(\mu\tau_m t - \frac{\mu}{2}\tau_m{}^2 + \omega_c\tau_m \right) \right). \quad (3)$$

After analog-to-digital conversion (ADC), the discrete time model of (3) satisfying Nyquist sampling can be derived by $y_k[n]=y_k(nT_s)$ for n=0,...,N-1.

III. WEIGHTED AVERAGE - ESPRIT

In this section, we show that the WA-ESPRIT is employed for the fine distance estimation of the received signal and that the proposed estimator combines the ESPRIT with a weighted average scheme by considering the signal-to-noise ratio (SNR) of the received signal.

A. ESPRIT

The ESPRIT can be represented as follows. Let $\mathbf{Y}=[y_1[n],..., y_K[n]]^T$ can be a set of snapshots from K antenna arrays. Then, autocorrelation matrix $\boldsymbol{R_k}$ can be expressed by

$$\boldsymbol{R} = \sum_{n=0}^{N-1} \boldsymbol{Y}\boldsymbol{Y}^H \quad (4)$$

The eigenvalue decomposition (EVD) of the autocorrelation matrix $\boldsymbol{R}$ has the form given by

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{S}_{N\times M} & \boldsymbol{G}_{N\times(N-M)} \end{bmatrix} \begin{bmatrix} \lambda_0 & & \\ & \ddots & \\ & & \lambda_{L-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{S}^*_{N\times M} \\ \boldsymbol{G}^*_{N\times(N-M)} \end{bmatrix} \quad (5)$$

where signal eigenvector matrix $\boldsymbol{S}=[s_0, \ldots, s_{M-1}]$ contains M eigenvectors which span the signal subspace of the correlation matrix, noise eigenvector matrix $\boldsymbol{G}=[g_0, \ldots, g_{N-M-1}]$ means N-M eigenvectors spanning the noise subspace of the correlation matrix and λ_n denotes a n-th eigenvalues of the correlation matrix. The largest M eigenvalue $\lambda_{0,\ldots,}\lambda_{M-1}$ correspond to the M eigenvectors of $\boldsymbol{S}$. The other eigenvalue $\lambda_M \ldots \lambda_{L-1}$ correspond to the eigenvectors of $\boldsymbol{G}$ such that $\lambda_M = \ldots = \lambda_{L-1}=\sigma^2$. Let us define $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ matrix, which is $\boldsymbol{S}_1=[\boldsymbol{I}_{m-1}\ \boldsymbol{0}]\boldsymbol{S}$ and $\boldsymbol{S}_2=[\boldsymbol{0}\ \boldsymbol{I}_{m-1}]\boldsymbol{S}$. The sub-matrices, which showed in [11] , are factorized by

$$\boldsymbol{S}_1 = \boldsymbol{A}_1\boldsymbol{C} \text{ and } \boldsymbol{S}_2 = \boldsymbol{A}_1\boldsymbol{D}\boldsymbol{C} = \boldsymbol{S}_1\phi \quad (6)$$

where $\boldsymbol{A}_1=[\boldsymbol{I}_{m-1}\ \boldsymbol{0}]\boldsymbol{A}$, $\boldsymbol{D}$=diag[$\delta_{0,\ldots,}\delta_{M-1}$], δ_m denotes the frequency of the transformed sinusoid for the m-th path (i.e. $\delta_m=\mu\tau_m T_s$), $\phi = \mathbf{C}^{-1}\mathbf{D}\mathbf{C}$ and $\boldsymbol{C}$ denotes the non-singular transformation matrix of M by M. So ϕ has the same eigenvalues as $\boldsymbol{D}$. ϕ is uniquely determined given by

$$\phi = (\boldsymbol{S_1}^*\boldsymbol{S_1})^{-1}\boldsymbol{S_1}^*\boldsymbol{S_2} \ . \quad (7)$$

Among the number of ϕ, the first angle estimate is found by

$$\hat{\phi} = \sin^{-1}\left(\frac{1}{\pi}\angle(v_1)\right) \quad (8)$$

where $\angle(\cdot)$ means the phase angles for a complex signal and v_1 are first eigenvalue of ϕ.

B. Weighted Average Scheme

From (3), when the channel is not varied in processing time, the WA-ESPRIT exactly estimated the angle frequency that is affected by noise. When the power of the received signal is applied to weighted average scheme, the proposed WA-ESPRIT has better performance.

To estimate a frequency that is affected by noise, a weighted average angle frequency is expressed using

$$\hat{\phi}_{WA} = \sum_{i=0}^{L-1} \hat{\eta}_i\hat{\phi}_i \quad (9)$$

where η_i denotes the signal power of i-th received signal, $\hat{\phi}_{wa}$ is the proposed angle estimation result.

IV. PERFORMANCE ANALYSIS OF THE WA-ESPRIT RADAR

This section analyzes the performance of the WA-ESPRIT radar. When the receiver is assumed to be perfectly received without unwanted frequency, the detection probability of the coherent receiver can be given by Equation (10) and (11) [5].

$$P_D = Q(Q(P_{FA}) - \sqrt{N\cdot SNR}) \quad (10)$$

$$Q = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt \quad (11)$$

where P_{FA} and P_D denote the false alarm rate and detection probability, respectively. In order to evaluate the detection probability of the proposed radar receiver, we assume that the FFT and the square block of received signal have same distribution characteristics as those of the square of the Gaussian random variable. If received signal represents noise

alone, then the probability density functions of noise can be calculated at the receiver as

$$p_0(y_k) = \frac{1}{\sigma^n 2^{n/2} \Gamma\left(\frac{1}{2}n\right)} y_k^{(n/2)-1} e^{-D/2\sigma^2} \quad (12)$$

where the *n* degree central chi-square distribution with zero mean and variance σ^2 and $\Gamma(n)$ is the gamma function. Then, when a received signal is existed with signal and noise, then the probability density functions of y_k can be calculated at the receiver output as

$$p_1(y_k) = \frac{1}{2\sigma^2}\left(\frac{y_k}{s^2}\right)^{(n-2)/4} e^{-(s^2+y_k)/2\sigma^2} I_{(n/2)-1}\left(\sqrt{y_k}\,\frac{s}{\sigma^2}\right) \quad (13)$$

where the n degree non-central chi-square distribution with s^2 mean and variance σ^2 and $I_\alpha(x)$ is the α th-order modified Bessel function of the first kind. The probability of false alarm, P_{fa}, is defined as the probability that a sample $y_k[n]$ will exceed the defined threshold when noise alone is present in the radar receiver,

$$P_{fa} = \int_T^\infty \frac{1}{\sigma^n 2^{n/2} \Gamma\left(\frac{1}{2}n\right)} y_k^{(n/2)-1} e^{-y_k/2\sigma^2} dy_k \quad (14)$$

where the n degree is the same as the *N*-point FFT and *T* is the defined threshold level. The probability of detection, P_D, is the probability that a sample $y_k[n]$will exceed the defined threshold in the case of noise plus signal in the radar receiver,

$$P_D = \int_T^\infty \frac{1}{2\sigma^2}\left(\frac{y_k}{s^2}\right)^{(n-2)/4} e^{-(s^2+y_k)/2\sigma^2} I_{(n/2)-1}\left(\sqrt{y_k}\,\frac{s}{\sigma^2}\right). \quad (15)$$

The relation between the detection probability and the false alarm rate of the squared non-coherent receiver is analyzed with *N*-point *FFT*, as in Eq. (12) such that

$$P_D = Q(Q^{-1}(P_{FA}/2) - \sqrt{N \cdot SNR}) + Q(Q^{-1}(P_{FA}/2) + \sqrt{N \cdot SNR}) \quad . \quad (16)$$

V. SIMULATION RESULTS

We present Monte-Carlo simulation results averaged over 10,000 estimates to evaluate the performance of the proposed algorithm. The angle estimation performance of the proposed algorithm is compared with that of the conventional algorithms such as ESPRIT-based angle estimation algorithm. This paper only takes into account the RMSE for single tone frequency. In the following simulations, we normally adopt the FMCW radar system with *M*=3 and *K*=4.

In Figure 2, the proposed algorithm is compared to other algorithms such as DFT and ESPRIT. Here, δ means the fractional number of the angle frequency. In case of *M* =3, the RMSE of the proposed algorithm has better performance than that of the ESPRIT at every angle frequency.

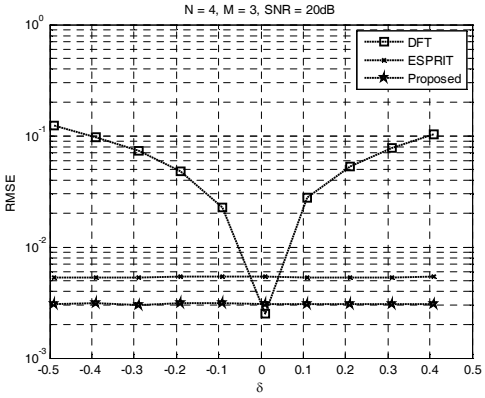


Figure 2. Performance comparison of the proposed method in various angle

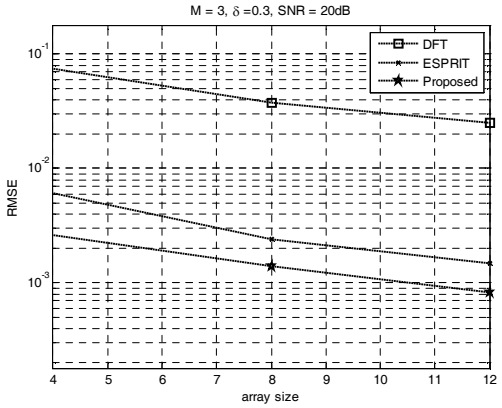


Figure 3. Performance comparison of the proposed algorithm for various antenna arrays *K*

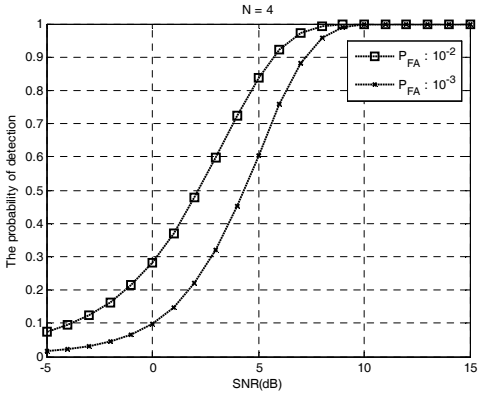


Figure 4. Performance comparison of the proposed algorithm for various false alarm rate

Figure 3 shows the RMSE of various estimators according to increasing of the sample N. For the change of the array size *K*, *K* increases from 4 to 12. When *K* increases, the RMSE of the proposed estimator improves. In particular, in the case of the proposed estimator, when *N* changes from 4 to 12, the RMSE characteristics improve by more than about 3 times, with a change from 2.59e-3 to 0.82e-3. Figure 4 shows the performance comparison of the proposed algorithm for various false alarm rates. After

SNR=10dB, the detection probability of P_{FA} =10^{-2} and 10^{-3} is same.

VI. EXPERIMENTS

In order to testify the effectiveness of the proposed estimation in a real environment, at Daegu-Gyeongbuk Institute of Science & Technology (DGIST) in Korea, we fulfilled various experiments in an anechoic chamber. We implemented a 24 GHz FMCW RF module which included a transmitting/receiving channel. The transmitter contained a voltage controlled oscillator (VCO), a frequency synthesizer, and a 26MHz oscillator used as the input of VCO. A frequency synthesizer controlled the input voltage of the VCO in order to generate the FMCW source. The source swept over the range of 24.05-24.25 GHz, i.e., a 200 MHz bandwidth. The receiver consisted of three LNAs, three mixers, three high-pass filters (HPFs) and three low-pass filters (LPFs). The receiver had an overall noise figure of 8 dB. The gain and noise figure of the LNAs were 14 dB and 2.5 dB, respectively. An RF signal was down-converted to an IF signal (beat signal) by the mixer.

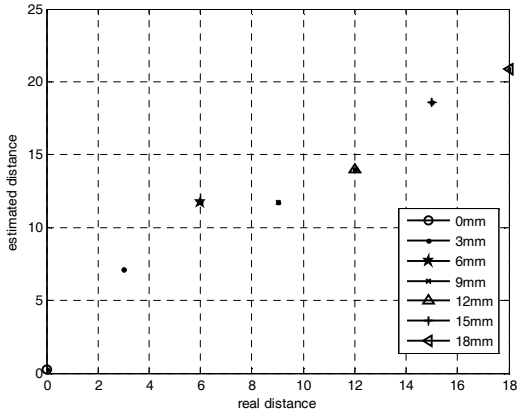


Figure 5. Experimental results

We verified the performance of the proposed method for various distance between radar and target in the chamber. When a target were placed at R=[0, 3, 6, 9, 12, 15, 18]mm for precise estimation in the anechoic chamber, respectively, the range map was derived as in Figure 5, where each colored dot indicates an estimation result. The proposed method can estimate the TOAs of various distances well.

VII. CONCLUSION

This paper proposed the fine angle frequency estimation system for radar-based WSN that takes into consideration the WA-ESPRIT. In order to improve the accuracy of the angle frequency estimation, the proposed algorithm applied a weighted average scheme according to the average signal amplitude of the received signal. We illustrate that the proposed estimator has better performance than the ESPRIT at M=3 and improves by more than about 3 times from K=4 to 2. In the future, we will make the various outdoor experiments.

ACKNOWLEDGMENT

This work was supported by the DGIST R&D Program of the Ministry of Science, ICT and Future Planning, Korea (15-RS-01) and the Ministry of Trade, Industry & Energy(MOTIE), Korea Institute for Advancement of Technology(KIAT) and DaeGyeong Institute for Regional Program Evaluation(DGIRPE) through the Leading Industry Development for Economic Region.

REFERENCES

[1] P. K. Dutta, A. K. Arora and S. B. Bibyk, “Towards radar-enabled sensor networks,” in Proc. 5th Int. Conf. Information Processing SensorNetworks (IPSN), Nashville, TN, Apr. 2006, pp. 467–474.

[2] M. S. Lee and Y. H. Kim, "Design and Performance of a 24-GHz Switch-Antenna Array FMCW Radar System for Automotive Applications," IEEE Transactions on Vehicular Technology, vol.59, no.5, pp.2290-2297, Jun 2010

[3] A. N. Lemma, A. J. Vanderveen and E.F. Deprettere, “Multiresolution ESPRIT Algorithm,” IEEE Transactions on Signal Processing, vol. 47, pp. 1722-1726, Jun. 1999.

[4] T. K. Sarkar and O. Pereira, “Using the matrix pencil method to estimate the parameters of a sum of complex exponentials,” IEEE Antennas Propagator Magazine, vol.37, pp.48–55, Feb. 1995.

[5] X. Li and K. Pahlavan, “Super-Resolution TOA Estimation With Diversity for Indoor Geolocation,” IEEE Trans. on Wireless Comm., vol. 3, pp. 224–234, Jan. 2004.

 ISBN: 978-1-61208-398-8

A Border-Oriented-Forward Routing Protocol for Large-Scale WSANs with Support to Actuator-Sensor-Actuator Communication

Luis Eduardo Lima, Juliana Garcia Cespedes, Mariá C. V. Nascimento and Valério Rosset
Science and Technology Department
Federal University of São Paulo - UNIFESP
São José dos Campos, São Paulo, Brazil
Email: {llima, jcespedes, mcv.nascimento, vrosset}@unifesp.br

***Abstract*—Wireless Sensor and Actuator Networks (WSANs) refer to a class of unattended wireless networks whose goal is to provide communication between distributed applications that perform the monitoring and control of certain characteristics of an environment. Among the vast number of WSANs applications, we focus on those designed to the natural disaster monitoring, such as forest fire control. It is therefore essential that the routing algorithms for these applications provide adequate scalability, reliability and energy efficiency. In addition, considering large-scale scenarios and the absence of direct communication between actuators, the message delivery reliability, fundamental for the mutual coordination of actuators, is a requirement not fully supported by the existing protocols. In order to cope with this shortcoming, in this paper we propose a novel Border-Oriented-Forward routing protocol (BOFP) for WSANs with support to actuator-sensor-actuator communication. We carried out simulations to attest the BOFP performance according to three metrics: the goodput, the overall end-to-end delivery delay and the energy consumption. The results of the simulations were statistically analyzed and suggested that BOFP satisfied the requirements of large-scale WSANs.**

Keywords–WSAN; Routing; Energy Efficiency; Reliability.

I. INTRODUCTION

Wireless Sensor and Actuator Networks (WSANs) have appeared as an extension of Wireless Sensor Networks (WSNs) in which the actuation tasks can be performed directly in a monitored environment. Such networks usually operate unattended and are composed by a set of heterogeneous devices mainly equipped with microprocessors, wireless communication adapters, sensors and actuation mechanisms. These devices may play particular roles in the field, such as data gathering that is performed by low-cost devices, named sensor nodes. In this case, the desired condition of the environment is maintained by more complex and expensive devices known as actuators nodes. A substantial number of WSANs applications can be enumerated, as, e.g., precision agriculture, natural disaster monitoring, tracking and location in hospital environments, home, industrial automation and so on [1].

In particular, applications intended to monitoring large geographical areas demand a large number of sensor nodes. In such cases, the actuators density might be much lower than the sensor nodes density, primarily due to the cost of actuator nodes. In large-scale unattended WSAN applications, e.g., the forest fire control or the landslide monitoring, the coordination between the actuator nodes can be considered essential [2]. Moreover, in these specific applications, a common assumption is to consider that actuators can communicate directly to each other. However, in such a scenario, the communications are subject to fail due to either obstacles or long distances between the actuators. Therefore, the network scalability is limited by the maximum communication range of the wireless adapter of the actuators. For this reason, routing protocols designed for large-scale WSANs must be scalable and also to provide both appropriate efficiency of energy consumption and adequate reliability on message delivery.

One can find several WSAN routing protocols in the literature that may be in line with some of the aforementioned requirements [3]–[13]. However, among them, the on-demand gradient-based routing protocol (DGR) [11] was specially designed to provide communication between actuators, named Actuator-Sensor-Actuator Communication (ASAC), by defining routing paths linking the set of sensor nodes. However, due to its on-demand feature, the DGR may not achieve satisfactory overall data delivery reliability and energy conservation required for some WSAN applications.

In line with this shortcoming, in this paper, we present a new routing protocol, named Border-oriented routing protocol (BOFP), specially designed to large-scale WSANs with support to ASAC communication. The BOFP is a gradient-based protocol that organizes the sensor nodes in levels and defines some of them as the relays of the actuators messages. Therefore, the BOFP automatically delimits broadcast regions in the network by defining special sensor nodes, located in border of the regions, each of them here named Border Node (BN). The BNs have as main task to establish routing paths between two or more actuator nodes. However, the BNs can also prevent the propagation of broadcast of messages (broadcast storm) outside the delimited regions, consequently, reducing the overall intra-network collision rate.

For assessing the performance of the BOFP, we carried out simulations considering different scenarios. Moreover, we compared the results of the BOFP with those from DGR with regard to their energy consumption efficiency, delay and message delivery reliability considering the event detection and ASAC data traffics. Consequently, the results achieved by the BOFP outperformed those achieved by the DGR protocol.

The remaining sections of this paper are organized as follows. In the next section, we briefly present the protocols closely related to the proposed strategy. Section III presents the BOFP specification and the WSAN model employed in this study. The performance evaluation of the BOFP and the results are presented in Section IV. To sum up, we conclude the paper by presenting some final remarks in Section V.

 ISBN: 978-1-61208-398-8

II. Related Work

Most of the existent protocols [3]–[6] [8]–[10] [12] [13], which have some relation to the objective of the present study, use a central controller to configure or coordinate the overall operation of the network. Consequently, these protocols have scalability limitations and may not be suitable for large-scale WSANs. Differently, the BOFP, proposed in this paper, makes use of a distributed algorithm in which the network operation is decentralized and the communication between sensors and the actuators occur locally, with sensors within their neighborhood. Besides, with exception of the study introduced in [8], the protocols earlier cited in this section do not consider ASAC. Additionally, in spite of using ASAC, the protocol proposed in [8] assumes the existence of a specific hardware for directional antennas and the knowledge of the geographic location of nodes both not regarded in this paper.

To the best of our knowledge and considering only the routing protocols designed for WSANs, the approach most related to this work is the DGR, proposed by Guo et al. [11].

The operation of the DGR starts with the dynamic gradient setup phase, where announcement messages (*ADV*) are propagated, hop-by-hop, from each actuator to all sensor nodes. In this phase, all sensor nodes in the network calculate a *k* value that will correspond to a cost gradient. Added to this *k*, an energy gradient *s* and a balance coefficient α define a backoff timer t_b that depending on its value, it reveals whether or not a sensor node will belong to a routing path. Accordingly, in the case that an actuator is supposed to perform an ASAC, and from it there does not exist a routing path, it initiates a routing path establishment phase. The actuator broadcasts a transmission-request message (*TR*) and waits for the response of the sensor nodes.

Each sensor node that received the TR calculates its t_b that is the required backoff timer for response. Since the t_b value is influenced by the residual energy of the sensor node, the node with the largest amount of residual energy and that is the nearest (has the smallest *k*) to the destination provides the lowest t_b and, consequently, answers first. The answer message is called transmission-agreement message (*TA*). The source of the *TA* message is then warned, by the actuator, that will become part of the routing path and does the same steps of the actuator, by broadcasting a *TR* to its neighbors, to find the next hop to the destination. This process is repeated until a *TR* reaches the destination. From this point on, data messages can be sent by the source actuator to the destination through the routing path found.

The DGR relies on the assumption of the existence of high traffic between the actuator nodes. Consequently, by just considering the actuator activities, the DGR protocol does not explicitly consider the influence of the data traffic generated by the sensor nodes for constructing the routing paths between the actuators. In the DGR, routing paths are built without taking into account the distance between the nodes involved in a point-to-point transmission. This can reduce the chances of transmission success leading to a decrease in the reliability of data delivery in large-scale networks.

Bearing all these limitations pointed up in the DGR in mind, we developed the BOFP.

III. Protocol Specification

In this paper, we introduce the BOFP that overcomes some limitations pointed up in the previous section. As well as the DGR, the BOFP builds routing paths for the communication between actuators through the set of sensor nodes. Unlike the DGR, in our proposal we assume that monitoring applications for event detection produce low traffic between actuators, and this traffic is active only when critical events are detected. Another difference between the BOFP and the DGR is with respect to the routing path establishment. In the DGR, routing paths are established on-demand while in the BOFP they are established only at the beginning of the protocol operation. Additionally, the BOFP uses a threshold of Received Signal Strength Indicator (RSSI) to determine the maximum distance that a node can have in relation to the transmitter, to make part of a routing path. Thereby, the BOFP allows the adjustment of the RSSI threshold in order to maintain the reliability of data delivery in adequate levels.

A. The WSAN model and Assumptions

In this paper, we consider a stationary and unattended WSAN composed by a set of actuator nodes and a set of sensor nodes unaware about their geographic location coordinates. The set of sensor nodes is homogeneous with regard to the hardware and software capabilities. Accordingly, the sensor nodes are equipped with the same radio device in which the transmission power is not dynamically adjustable. All actuator nodes are homogeneous in hardware and software capabilities and use the same transmission power of sensor nodes for short range communication. Although actuator nodes are assumed to be more powerful than sensor nodes, we consider that, due to the long distances implied in large-scale WSANs, any of them does not directly transmit messages to other actuator.

Nevertheless, for the source and the destination identification, we assume that every node (sensor or actuator) is assigned to an unique identity, as MAC address. Additionally, each sensor node keeps a routing table where each entry indicates the next hop and the distance (in hops) to a specific actuator node. Thus, we consider that each sensor node will always send data messages to the closest actuator node, related to the entry in the routing table with the smallest distance to it. Finally, we also consider that all nodes do not need to be clock synchronized. Taking this WSAN model into account, we specify the introduced protocol in the next section.

B. The BOFP General Operation

The proposed protocol has a two-phase approach consisting of a startup phase (S-Phase) and a communication phase (C-Phase). The S-Phase comprises the execution of two asynchronous distributed procedures with the purpose of discovering routing paths from sensor nodes to actuators and from actuators to actuators.

The first procedure, here called Node-to-Actuator Discovery Path Procedure (N2A-DP), is executed at the beginning of the S-Phase. In this procedure, each actuator node broadcasts a Route Discovery Message (RDM) to all sensor nodes in its communication range. The *RDM* carries a tuple of three integer values: the level counter (*lc*) initially set to zero, the unique source sensor node identity (*snid*) and the actuator unique identity (*aid*). All RDM sent by actuator nodes have the *snid*

 ISBN: 978-1-61208-398-8

set to zero. At this point, in the S-Phase, all sensor nodes execute the Algorithm 1.

It is important to notice that, according to the proposed algorithm, a given sensor node accepts a received message if and only if the Signal Strength Indicator (RSSI) of the received message, $m.rssi$, is above a specific threshold value γ . The value for γ is considered variable and defined according to the application requirements. After receiving and accepting a RDM, say *m*, a sensor node verifies whether or not the received *aid*, $m.aid$, matches some *aid* in the routing table entry set (*RT*). In the first case, if the received *lc* value is smaller than the *lc* value stored in the *RT*, it updates its *RT* entry related to the *aid* with the received values of *lc* and *snid*. In the second case, the sensor node includes the received values in the *RT*.

Afterwards, each sensor node, receiver of *m*, generates and broadcasts a replica of *m*, say *m'*, which carries the following information: a *lc'* with the value of *lc* increased by one; and both the *aid* and the *snid* are set as the sensor node own identity. This process is repeated by every sensor node receiver of *m* or its replicas. Therefore, the *lc* stored in any entry of *RT* indicates how far (in hops) the sensor node is from a given actuator node.

Notwithstanding, regarding to distinct *lc* values of *aid* entries stored in the *RT*, $RT.aid.lc$, a sensor node may find similar values of the received *lc*, $m.lc$.Here we define similar as those $RT.aid.lc$ that differ at most by one to the value of the received *lc*. Whenever a sensor node finds similar values of a received *lc* in the *RT*, it becomes a border node (BN) of the actuator of the received *lc* and of every distinct actuator whose *lc* is similar in the RT. As a matter of fact, the following steps with regard to the BN happens for every actuator whose *lc* is similar to the received *lc*. However, the operations are pairwise, being one of the pairs always the source of the received *lc*. By considering one of these pairs of actuators (represented by the received *aid* and the *aid* stored in *RT*), from the point it is defined as BN on, the sensor node stops the RDM propagation and starts a complementary Actuator-to-Actuator path Discovery Procedure (A2A-DP).

In the A2A-DP, each BN sends two Actuator-to-actuator Route Discovery Messages (ARDMs) via unicast to the pairs of actuators ($m.aid, RT.aid$). On the one hand, the ARDM for the actuator $m.aid$ contains the information of the distance in hops between the BN and the actuator $RT.aid$. On the other hand, the ARDM for the actuator $RT.aid$ contains the information regarding the distance in hops between the BN and the actuator $m.aid$. The ARDM has a very similar structure to the RDM. The sensor nodes in the path from the BN to the actuator $m.aid$ (or $RT.aid$) will handle the ARDM as they did with the RDM by including in their RT the distance and the next hop to the actuator $RT.aid$ (or $m.aid$) and by increasing the values of *lc* in the ARDM for every retransmission. Immediately after receiving the ARDM, both actuators store the next hop and the distance to each other in their *RT*. It is worth mentioning that, for a given path between two actuators, there will be always only one BN. Considering a path between two actuators whose number of sensor nodes is even, two sensor nodes are candidate to be BN. But, is this case, the first candidate sensor node to become BN will stop the propagation of the RDM. Consequently, the second candidate sensor node will not receive the corresponding RDM.

Algorithm 1: Algorithm for S-Phase executed in sensor nodes

```
input: A received message m;
       The routing table entry set, RT;
       The sensor node state, BN, initially set to FALSE;
       The ID of the sensor node, myid.
if m.rssi ≥ γ then
    if m is RDM then
        if m.aid ∈ RT then
            if m.lc < RT.(m.aid).lc then
                Updates the m.lc, m.snid values of
                RT.aid entry;
            end
        else
            Includes a new entry in RT with m.aid, m.lc
            and m.snid values;
        end
        for each RT.aid ≠ of m.aid do
            if m.lc = RT.aid.lc or [ m.lc − RT.aid.lc] =1
            then
                set BN to TRUE;
                set m' to ARDM with (aid=m.aid,
                lc=m.lc+1, nexthop=RT.aid.snid);
                set m'' to ARDM with (aid=RT.aid,
                lc=RT.aid.lc, nexthop=m.snid);
                sendbyUnicast (m', m'');
            end
        end
        if not BN then
            set m' to RDM with (aid=m.aid,
            lc=m.lc+1);
            Broadcast(m');
        end
    else
        if m is ARDM then
            if m.snid = myid then
                if m.aid ∉ RT then
                    Include m.aid in RT with m.lc,
                    m.snid values;
                    for each aid ∈ RT ≠ of m.aid do
                        set m' to ARDM with
                        (aid=m.aid, lc=m.lc+1,
                        nexthop=RT.aid.snid);
                        sendbyUnicast(m');
                    end
                end
            end
        end
    end
end
```

Finally, in the C-Phase, sensor nodes and actuators are able to send data messages. In the C-phase we consider two types of unicast data messages: the Sensor-to-actuator Data Message (SDM) and the Actuator-to-actuator Data Message (ADM). As stated before, the SDMs are transmitted, from sensor nodes, through the shortest path to their nearest actuator. Differently, the ADMs include the *aids* of both source and destination actuators.

IV. Performance Evaluation

For evaluating the performance of the BOFP, we propose three metrics: end-to-end delay, energy consumption and good-put. The main reason behind the use of these specific metrics

 ISBN: 978-1-61208-398-8

is the overall assessment of the proposed protocol mainly by means of energy conservation efficiency and reliability on message delivery. Each of these assessment measures, simulation parameters, scenarios and performance evaluation results are gone into detail in the next sections.

A. Simulation Models, Scenarios and Parameters

We compared the performances of the BOFP, its variations, and the DGR into three extents of square sensing areas, representing small, moderate and large-scaled scenarios, as we can observe in the summary of the simulation parameters indicated in Tables I and II. Additionally, for each sensing area we set different numbers of nodes. We also consider that the application periodically senses the environment, at every five seconds, and depending on the simulation purpose, it may generate one data message whenever it detects an event.

Table I summarizes the parameters for the first simulation model designed to assess the performance of both protocols with respect to the event detection. We configured two static events to be activated in two non simultaneous periods of 50 seconds during the execution of the simulation. The sensor nodes were uniformly placed in the field while the actuator nodes were arbitrarily placed out of the communication range of each other. We set the sensors nodes in the DGR to send messages to their closest actuator. To better assess the impact of the main differential of BOFP, the BN nodes, in addition to the DGR Protocol, we also modeled simplified version of the BOFP, here named Simple Gradient Protocol (SGP). This simplified version of the BOFP has the same procedures developed in the BOFP, however, without the BNs. Therefore, the SGP uses the same implementation of the gradient based level count as well as all the discovery procedures used in the BOFP without the limitation of broadcast messages done by the BN nodes. We also include the RSSI threshold in the SGP implementation in order to compare s performance with the BOFP. The RSSI values were defined by analyzing, for each scenario, whose values maximized the delivery ratio in point-to-point communication.

Table II summarizes the parameters of the second simulation model designed to compare the performance of the BOFP with the DGR with respect to the ASAC. In this second model, we do not consider the traffic generated by sensor nodes as result of event detection, as defined in [11]. Differently of the former model, we deployed only two actuator nodes in the opposite extremity sides of the sensing field. We also consider that the traffic load is generated by one of the actuators, that sends one message at every five seconds.

We designed and executed the simulation models in OMNeT++/Castalia environment [14]. For each simulation scenario, the results we report correspond to the average of 40 independent executions with different seeds. Moreover, since the samples did not follow a normal distribution we proceeded with the statistical analysis, considering each scenario independently, by applying the Kruskal-Wallis test [15].

B. Metrics

1) End-to-end delay: Here, the end-to-end delay is the elapsed time between the message generation by any sensor node until the delivery of such message to an actuator.

TABLE I. PARAMETERS OF THE MODEL SETUP FOR TESTING SCENARIOS OF SENSOR-ACTUATOR COMMUNICATION.

	Small Scale	Moderate Scale	Large Scale
Event1:(x,y) coord.	50,50	50,50	50,50
Event2:(x,y) coord.	170,170	900,900	1900,1900
Event Detec. radius(m)	36	36	36
Area (mxm)	350x350	1000x1000	2000x2000
Sensor nodes (unit)	100	400	1600
Actuator nodes (unit)	4	5	8
Power Radio (dBm)	10	10	10
Model Radio	CC1000	CC1000	CC1000
Sampling interval (s)	5	5	5
Events (unit)	2	2	2
Simulation time (s)	200	200	200
Executions (unit)	40	40	40
Message size (bytes)	512	512	512
Transmission rate (kbps)	19,2	19,2	19,2
γ values (dBm)	-86	-92	-92

TABLE II. PARAMETERS OF THE MODEL SETUP FOR TESTING SCENARIOS OF ACTUATOR-ACTUATOR COMMUNICATION.

	Small Scale	Moderate Scale	Large Scale
Area (mxm)	350x350	1000x1000	2000x2000
Sensor nodes (unit)	100	400	1600
Actuator nodes (unit)	2	2	2
Power Radio (dBm)	10	10	10
Model Radio	CC1000	CC1000	CC1000
Sampling interval (s)	5	5	5
Simulation time (s)	270	270	270
Executions (unit)	40	40	40
Message size (bytes)	512	512	512
Transmission rate (kbps)	19,2	19,2	19,2
γ values (dBm)	-86	-92	-92

2) Energy Consumption: The energy consumption is a very important metric to assess the protocol efficiency. We estimate the individual energy consumption for a single transmission (E_{Tx}) and for a reception (E_{Rx}) considering the energy model proposed by Heinzelman et al in [16]:

$$E_{Tx}(k,d) = E_{elec} * k + e_{amp} * k * d^2 \quad (1)$$

$$E_{Rx}(k) = E_{elec} * k \quad (2)$$

where k is the number of bits to be transmitted with distance d, considering the energy spent $E_{elec} = 50nJ/bit$ for both transmission/reception and antenna amplification $e_{amp} = 100pJ/bit/m^2$. The total energy consumption of the network is given by the sum of the energy spent in each transmission/reception as given by the Equations 1 and 2 [16].

3) Goodput: This measure, highly related to the message delivery reliability, consists in the ratio of the total number of original application data messages to the total number of messages delivered to the actuators nodes.

C. Simulation Results and Discussion

1) Attesting the benefits of BN: The charts depicted in Figure 1 present the results corresponding to the scenarios of the first simulation model designed to evaluate the performance of the protocols considering the event data traffic generated by sensor nodes. In the chart, the black dots represent the mean values whereas the letters *a*, *b* and *c* indicate the statistical equivalence of the samples. For example, given the results of a specific metric for each distinct scenario, if two boxes are with the same letter, this means that they are not, according to the Kruskal-Wallis test, statistically different regarding this metric. Otherwise, they have a significant difference. One may noticed that, for the trhee metrics, BOFP behaves slightly better than the others as the extent of the scenarios increases. The statistical analysis has showed that the results obtained by BOFP and SGP are similar considering the goodput. However, the reduced energy consumption of BOFP compensates, since it can deliver almost the same average number of messages as SGP while requiring less energy. Additionally, BOFP presents an overall end-to-end delay smaller than other protocols.

2) Performance in ASAC: The main goal of this evaluation is to compare the performance of the BOFP with the DGR considering the ASAC. As indicated in Figure 2, it is noteworthy that the BOFP outperforms the DGR with respect to the goodput and the end-to-end delay. Consequently, reliability with regard to data delivery of BOFP is better than from DGR. Moreover, the DGR has a better performance than the BOFP when considering the energy consumption. The reason is the number of transmissions executed by the BOFP: as the BOFP transmits much more messages than DGR, the energy consumption of the BOFP is higher. Finally, we also observed that the performance of the BOFP is significantly higher than of DGR because the latter is sensible to communication faults primarily during the routing path establishment phase execution.

3) Scalability properties of BOFP: In order to attest the scalability of the proposed protocol, we compared the goodput samples considering the three distinct extent scenarios. Figure 3 presents the results of this analysis. As stated before, distinct letters represent the statistical difference between the samples. According to this analysis, we did not find significant differences in the results. Thus, BOFP may cope with scalability requirement for large scale WSANs.

V. Conclusion and Future Works

In this paper, we addressed the problem of performing ASAC in large-scale WSANs. Additionally, for this type of WSAN, the protocols found in the literature revealed a shortcoming of not provide the adequate levels of energy efficiency and reliable data delivery. In line with this shortcoming, in this paper, we proposed a new routing protocol, the BOFP, specially designed for large-scale WSANs. Accordingly, for assessing the performance of the BOFP, we carried out simulations in different scenarios considering both data traffics of event detection and the ASAC. By comparing the BOFP with other protocols we observed results which revealed the benefits of the BN approach in determining the routing paths. The results also showed how the BOFP may achieve adequate reliable data delivery levels without compromise the energy efficiency. As future research, we intend to define a path reestablishment procedure in order to evaluate the BOFP considering the presence of multiple mobile actuators as well as to estimate the overall network lifetime. Additionally, we intend to proceed with the experimental analysis to evaluate BOFP performance in a real environment.

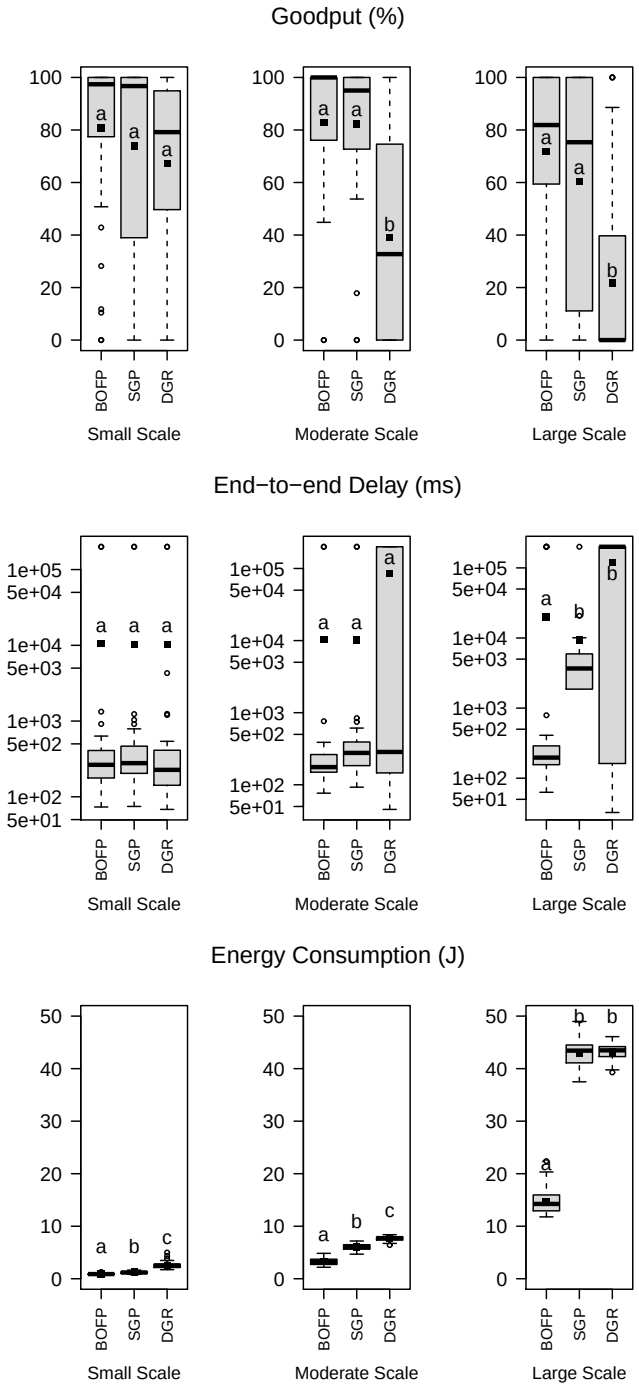


Figure 1. Results of Goodput, End-to-end Delay and Energy Consumption considering the event data traffic generated by sensor nodes.

Acknowledgment

The authors thank to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for its financial support.

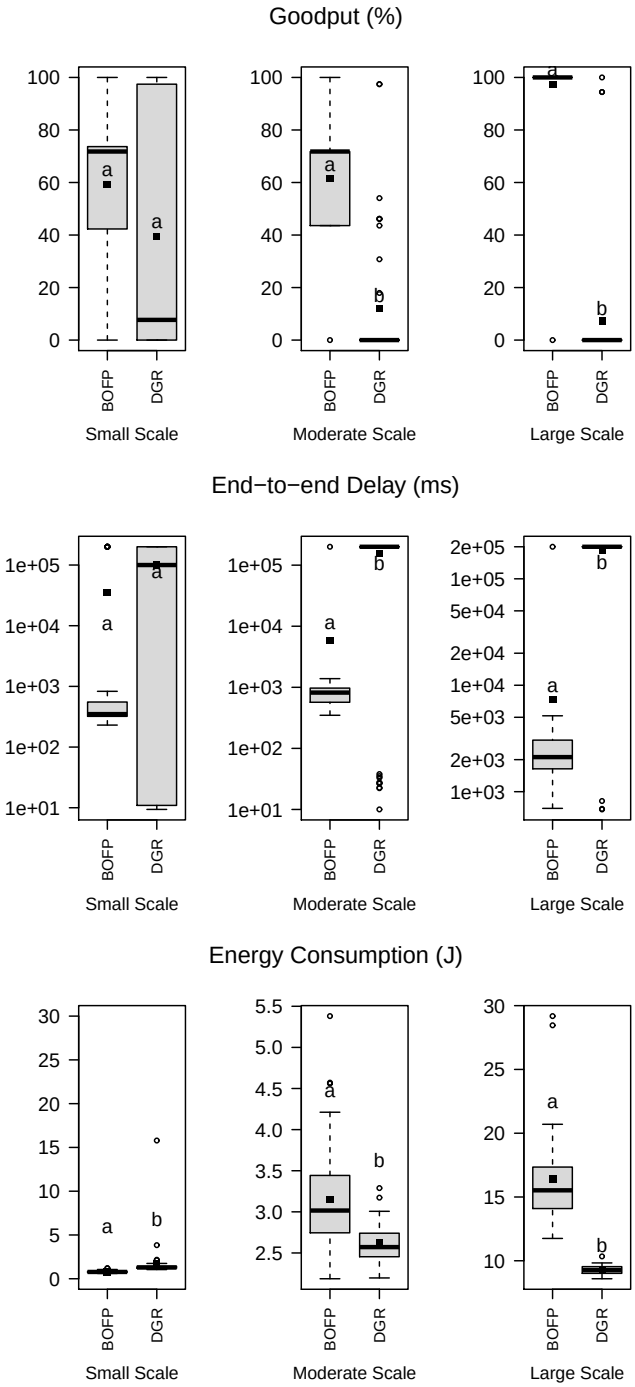


Figure 2. Results of Goodput, End-to-end Delay and Energy Consumption considering ASAC data traffic generated by one actuator.

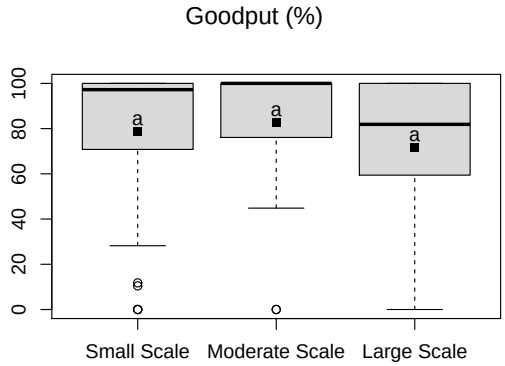


Figure 3. Comparison between the Goodput results, achieved by BOFP, considering the three scenarios extent.

REFERENCES

[1] I. F. Akyildiz and I. H. Kasimoglu, "Wireless sensor and actor networks: research challenges," Ad Hoc Networks, vol. 2, no. 4, 2004, pp. 351 – 367.

[2] H. Salarian, K.-W. Chin, and F. Naghdy, "Coordination in wireless sensor-actuator networks: A survey," Journal of Parallel Distributed Computing, vol. 72, 2012, pp. 856–867.

[3] W. Hu, N. Bulusu, and S. Jha, "A communication paradigm for hybrid sensor/actuator networks," in 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2004), vol. 1, 5-8, September 2004, pp. 201–205.

[4] E. Cayirci, T. Coplu, and O. Emiroglu, "Power aware many to many routing in wireless sensor and actuator networks," in Proceedings of the Second European Workshop on Wireless Sensor Networks, Jan-2 Feb 2005, pp. 236 – 245.

[5] H. Peng, W. Huafeng, M. Dilin, and G. Chuanshan, "Elrs: an energy-efficient layered routing scheme for wireless sensor and actor networks," in 20th International Conference on Advanced Information Networking and Applications, AINA 2006., vol. 2, April 2006, pp. 452 – 460.

[6] A. Boukerche, R. B. Araujo, and L. Villas, "A wireless actor and sensor networks qos-aware routing protocol for the emergency preparedness class of applications," in Proceedings of the 31st IEEE Conference on Local Computer Networks, November 2006, pp. 832–839.

[7] ——, "Optimal route selection for highly dynamic wireless sensor and actor networks environment," in Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems, ser. MSWiM '07, vol. 0. New York, NY, USA: ACM, 2007, pp. 21–27.

[8] K. Selvaradjou, N. Handigol, A. Franklin, and C. Murthy, "Energy-efficient directional routing between partitioned actors in wireless sensor and actor networks," Communications, IET, vol. 4, no. 1, 5 2010, pp. 102 –115.

[9] M. Akbas and D. Turgut, "Lightweight routing with qos support in wireless sensor and actor networks," in Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, December 2010, pp. 1 –5.

[10] M. I. Akbas, M. R. Brust, and D. Turgut, "Sofrop: Self-organizing and fair routing protocol for wireless networks with mobile sensors and stationary actors," Computer Communications, vol. 34, no. 18, 2011, pp. 2135 – 2146.

[11] Y. Guo, Z.-z. Xu, C.-l. Chen, and X.-p. Guan, "Dgr: dynamic gradient-based routing protocol for unbalanced and persistent data transmission in wireless sensor and actor networks," Journal of Zhejiang University SCIENCE C, vol. 12, 2011, pp. 273–279.

[12] Z. Li and H. Shen, "A kautz-based real-time and energy-efficient wireless sensor and actuator network," in IEEE 32nd International Conference on Distributed Computing Systems (ICDCS), june 2012, pp. 62 –71.

[13] E. Cañete, M. Díaz, L. Llopis, and B. Rubio, "Hero: A hierarchical, efficient and reliable routing protocol for wireless sensor and actor networks," Computer Communications, vol. 35, no. 11, 2012, pp. 1392 – 1409.

[14] A. Boulis, "Castalia, a simulator for wireless sensor networks and body area networks. user's manual." NICTA, March 2011, last accessed 27-February-2015. [Online]. Available: http://castalia.npc.nicta.com.au

[15] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," Journal of the American Statistical Association, vol. 47, no. 260, 1952, pp. 583–621.

[16] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-eficient communication protocol for wireless microsensor networks," in Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS 00), January 2000.

 ISBN: 978-1-61208-398-8

A Study of the Effects of Electromagnetic Fields on Digital Television Antenna Radiation: A Simulation and Evaluation of Exposure

Diogo Seiji Ishimori
Computer and Telecommunication Engineer Faculty
Federal University of Pará, UFPA,
Belém, Brasil
e-mail: diogoishimori@gmail.com

Jasmine Araújo
Computer and Telecommunication Engineer Faculty
Federal University of Pará, UFPA, Bolsita CNPq-Brasil,
Belém, Brasil
e-mail: jasmine.araujo@gmail.com

Ramz Luís Fraiha Lopes
Computer and Telecommunication Engineer Faculty
Federal University of Pará, UFPA,
Belém, Brasil
e-mail: ramzfraiha@ufpa.br

Josiane Rodrigues
Computer and Telecommunication Engineer Faculty
Instituto de Estudos Superiores do Pará – IESAM,
Belém, Brasil
e-mail: josi@ufpa.br

Rita de Cássia Souza
Computer and Telecommunication Engineer Faculty
Federal University of Pará, UFPA,
Belém, Brasil
e-mail: ritasosil@gmail.com

Gervásio Cavalcante
Computer and Telecommunication Engineer Faculty
Federal University of Pará, UFPA,
Belém, Brasil
e-mail: gervasio@ufpa.br

Abstract—**Electromagnetic compatibility is the ability of systems or equipment to be tested in the intended environment through levels of efficiency without degradation caused by electromagnetic interactions. Compliance with safety limits requires tests to measure the device when operated at w maximum power maintained over an average period of time. However, the average transmitted power of many of these devices depends on a range of system parameters, such as power control and others. Several agencies now require these devices to be subjected to safety limit tests, but are faced with the problem of a lack of standardized assessment procedures. This is a cause of anxiety among agencies, industry and consumers. The objective of this study is to examine procedures that can determine the maximum permissible exposure to radiation and that are scientifically and technically sound. Compliance with standards can be investigated on the basis of Specific Absorption Rate (SAR) measurements and derived quantities, i.e, the electric field and magnetic field. In this study, compliance with the electric field was carried out through the development of simulations, to aid in the evaluation of the exposure of populations that currently are (or will be) the subject of digital TV services in the frequency range of 599 MHz.**

Keywords-electromagnetic compatibility; simulation; method of moments; antenna array

I. INTRODUCTION

Electromagnetic fields naturally come from three main sources: the sun (130 mW/cm^2 at all frequencies), storm activities (electric fields in the range of many volts per meter) and the Earth's magnetic field (in the order of 40 A/m). In the last hundred years, radiofrequency fields generated by man, with much higher intensities and with a very different spectral power distribution, have changed these natural electromagnetic fields; and as a result they are now under study. The radiofrequency fields are classified as non-ionizing radiation because the frequency is too low for the photon energy to ionize atoms but at a sufficiently high power density to heat body tissue.

The general public has been exposed to these electromagnetic fields since the mobile phone market is one of the fastest growing services in the telecommunications industry. Owing to the damage that can be caused by Non-Ionizing Radiation (NIR) emitted by base stations, the use of mobile phone stations has begun to be questioned and raised serious concerns regarding the adverse effects of radiation. This problem is not limited to mobile phones, but any antenna in the range of non- ionizing radiation.

Since 1974, many countries have conducted research to enable them to lay down safety standards for protection. The Environmental Health Division of the World Health Organization (WHO) and International Radiation Protection Association/ International Non-Ionizing Radiation Commitee (IRPA/INIRC) jointly drew up a number of reports that form a part of the program of the WHO criteria for Environmental Health, sponsored by the United Nations Environmental Programme (UNEP). Each report includes an overview of physical and technical features such as measurement, instrumentation and sources of NIR applications. These provide criteria for fixing the exposure limits in scientific experiments and adopting procedures with the NIR data. The guidelines complied with the International Commission on Non-Ionizing Radiation Protection (ICNIRP), set up in 1994, and stipulated the limits of exposure to electromagnetic

 ISBN: 978-1-61208-398-8

fields at frequencies between 9 kHz and 300 GHz. In Brazil, Resolution No. 303/2002 of the National Telecommunications Agency (ANATEL) National Council adopted the ICNIRP guidelines for limiting exposure to electromagnetic fields [1].

In [2], an estimate was made of the probable source of maximum exposure to electromagnetic fields from a radio-communication station, based only on information about the height of the antenna, half-power angle and tilt. In [3], there is an evaluation of the levels of non-ionizing radiation subjected to by people inhabiting an area adjacent to the Radio Base Station (RBS). These measurements show the relationship between power versus distance used to calculate the far field (the electric field in the region). This field is used to calculate the levels of exposure to non-ionizing radiation that the general public is exposed to. In [4], a computational method is proposed for predicting electromagnetic wave propagation in the UHF range using the Method of Moments (MoM). The authors employed a prediction model. This also estimates the protection zone from the frequency and feed power values. In [5] and [6], there is a study of the radiation level from mobile telephony base-station antennas using electromagnetic (EM) field simulation and making comparisons with the on-site EM field measurement. In [7], a minimum distance requirement based on a SAR simulation is proposed for mobile base station antennas, and in [8], the impact of Wi-Fi access points is investigated to determine electromagnetic field exposure. The evaluations were carried out through measurements and the simulation through a ray tracing method. Stratakis et al. [9] evaluated the authors evaluated the electromagnetic field exposure, using measurements generated by Wimax Base Stations.

In this study, methods were employed to assess the compliance of wireless systems with the established limits of electromagnetic exposure in humans. A method is proposed for numerical analysis. Exposure assessments of wireless systems, under a limited set of operating conditions, were performed to estimate the maximum levels of the electric field.

In this paper, a system is proposed and implemented that involves the calculation of the limits of human exposure, by assessing the electric field that is simulated by antenna arrays in the UHF Range. The electric field was obtained by numerical simulation through the software implementation of the Method of Moments, expanded from the dipole antenna (as described in [4]) to the antenna array. The results obtained can aid the regulatory agencies by providing information about the location, antenna heights or Effective Isotropic Radiated Power (EIRP) irradiated by these antennas.

This work is structured as follows: Section 2 will address the question of exposure levels. In Section 3, the tool developed to assess exposure is described. In Section 4, the results are shown. Finally, Section 5 summarizes the conclusions of the proposed work.

II. EXPOSURE LEVELS

The board of directors of the ANATEL Council as its meeting of July 15, 1999, decided to adopt, (as a provisional measure), the radiation limits proposed by the ICNIRP. These were based on an evaluative study of human exposure to radiofrequency electromagnetic fields from transmission stations of telecommunications services. The purpose of the ANATEL regulatory requirements is to set limits and define the evaluation methods and procedures that must be followed by radio stations when granted licences. In the case of human exposure to electric, magnetic and electromagnetic fields radio frequencies have to be in the range of 9 kHz to 300 GHz [4].

The parameters used to define exposure limits are the electric field, magnetic field and plane-wave equivalent power density, subject to basic restrictions. The basic restrictions are based on the health risks caused by exposure to electric fields, or magnetic and electromagnetic variables in time. Depending on the frequency of the field, the physical quantities used to specify these restrictions are current density (J), SAR and power density (S).

These limits correspond to those of the ICNIRP guidelines and have been established by employing quantities that can be more easily measured or calculated than the basic restrictions.

Table I shows the exposure limits of the general public to electromagnetic fields in a range of frequencies between 9 KHz and 300 GHz.

TABLE I. LIMITS OF THE GENERAL PUBLIC TO ELECTROMAGNETIC FIELDS IN A RANGE OF FREQUENCIES BETWEEN 9 KHZ AND 300 GHZ [10]

Radio Frequency Range	E Field Intensity (V/m)	H Range Intensity (A/m)	Power Density Seq (W/m^2)
9 KHz to 65 KHz	87	5	_____
0.065 KHz to 1 MHz	87	0.73 / f	_____
1 MHz to 10 MHz	$87 / f^{1/2}$	0.73 / f	_____
10 MHz to 400 MHz	28	0.073	2
400 MHz to 2000 MHz	$1.375\ f^{1/2}$	$0.0037\ f^{1/2}$	f / 200
2 GHz to 300 GHz	61	0.16	10

III. TOOL FOR MODELLING OF ELECTROMAGNETIC WAVES

Guerreiro et al. [4] proposed a tool that employed the method of moments to model a radiated field with a dipole antenna.

This study used a customization of a Matlab ® package described in [11][12] and extended the work [4] to predict the irradiation of an array of antennas. It is necessary to rely on Rao-Wilton-Glisson (RWG) basic functions [11], the electric field integral equation, and the feedingedge model of

 ISBN: 978-1-61208-398-8

the underlying MoM code [11] to stimulate radiation and the scattering of basic Radiofrequency (RF) waves, wireless communication antennas and microwave structures.

The customization implemented in this work is able to calculate the exposure level in a diagram based on the voltage signal in the array antenna feed, the antenna structure and the operation frequency, as input of the updated software. The flowchart of the code execution and the new software proposed, is shown in Fig. 1.

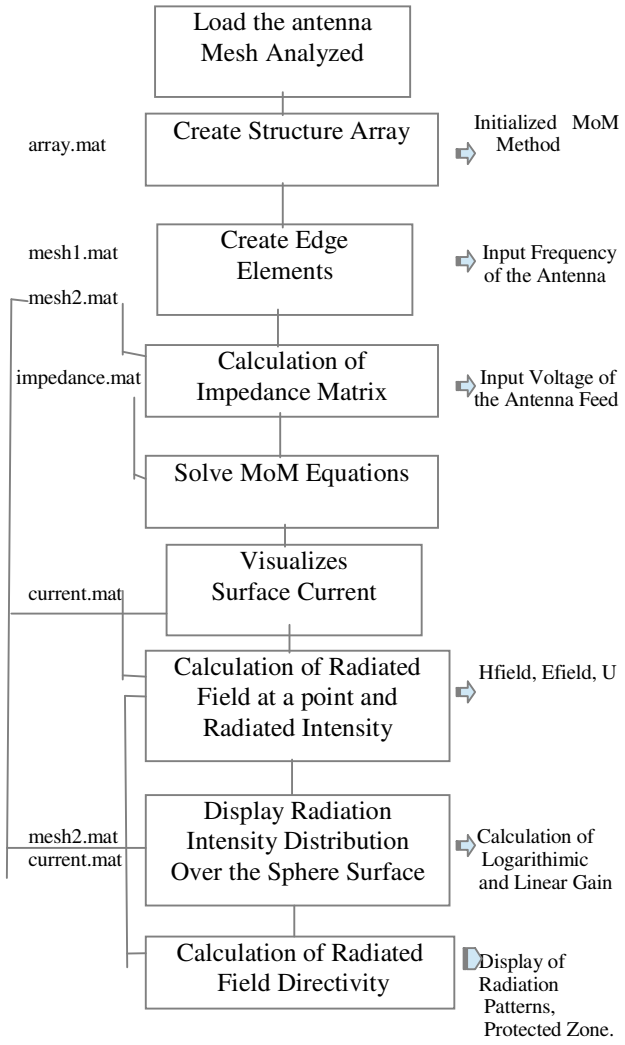


Figure 1. Flowchart of code execution sequence, adapted from [4], [11] and [12].

In the box 1, the antenna mesh is the input of the box 2 where the software builds the structure of the array and generates the file array.mat. This file, array.mat, is the input of the third box and where the edge elements are created generating the file mesh2.mat. The fourth box is responsible to calculate the impedance matrix and saving the file as impedance.mat. This file, impedance.mat, is the input of the MoM solver, after the surface current can be plotted using the sixth box and a file current.mat is generated. This last file is the input of the seventh box which calculates the radiated field in a single point. The box eighth and nineth are responsible to display the radiated field over a sphere surface and to calculate the radiated field directivity. Our customization was made in the box seventh to nineth where the evaluation of safe field according to ANATEL´s rules was implemented and finally a determination of the protected zone around the antenna can be defined according to the results described in the next section and showed in Fig. 2.

This tool mainly contributed to calculate the electric field in a practical way, aiding other possible works which needs this metric to develop proposed reseachers models.

Moreover, despite in [4], where the electric field value was calculated to dipole antenna. At this work the electric field is calculated to antenna array. This way, adding a new realistic possibility to attend wireless communications in the UHF range as digital TV.

IV. RESULTS

An array of antennas was used for the evaluation operating at a frequency of 599 MHz, horizontal polarization, with feed power of 1,333 kV, 0.288 m long and 0.5 cm wide. The antenna had a height of 125 meters. The antenna array used was of the broadside type with 8 elements and 0.25 cm spacing between the elements. These settings were adjusted in the multilinear.m file, dipolo599.mat, rwg3.m, rwg4.m and efield3.m [11].

In the MATLAB code efield1.m [11], the radiated scattered/field is calculated at a point that is outside the antenna surface. This corresponds to Box Seven in Fig. 1. The Efield1.m code was modified to discover where the electric field is considered to be non-compliant (larger than the minimum field permitted by the ANATEL regulations) and where it is considered to be safe.

In the code efield3.m [11], a customization was carried out to meet the objectives of this work, i.e., to trace a diagram exposure level around the antenna array in accordance with ANATEL regulations. Thus, a loop that can vary the radius of 1 meter to any value specified by the decision maker, was added to the code. The results of the electric field, the corresponding radius, and angle were stored for further analysis and are shown in Fig. 2.

The straight line described in the caption in Fig. 2 is the protection zone; this defines the contour of the protection zone. Electric field values below the ANATEL reference-point are represented by the plus (‘+’) signal and values above the limit of ANATEL are represented by a dot (‘.’) signal. In Fig. 2, it can be seen that the lowest height of a building near the antenna should be approximately 120 meters and the antenna distance should be about 20 meters. At a distance of about 40 meters from the antenna, any building could be built at about 125 meters of height, as shown in Fig. 2. ANATEL defines a range of 599MHz, in accordance with Table I, which is an exposure below 33.65 V/m.

 ISBN: 978-1-61208-398-8

It was confirmed that the MoM gives feasible results as shown in [13] even when it is employed to predict an array of antennas.

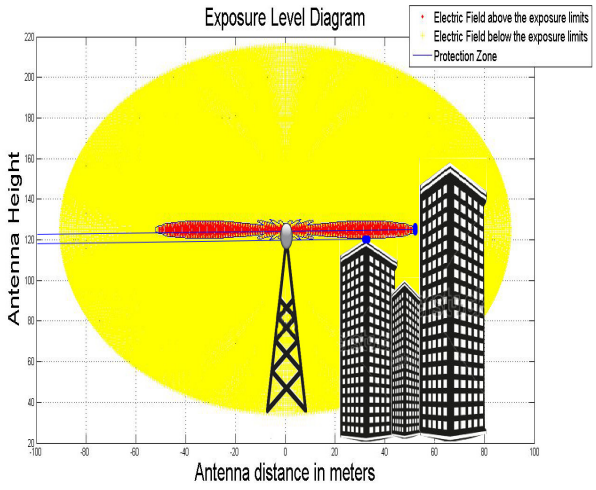


Figure 2. Diagram of Exposure Levels

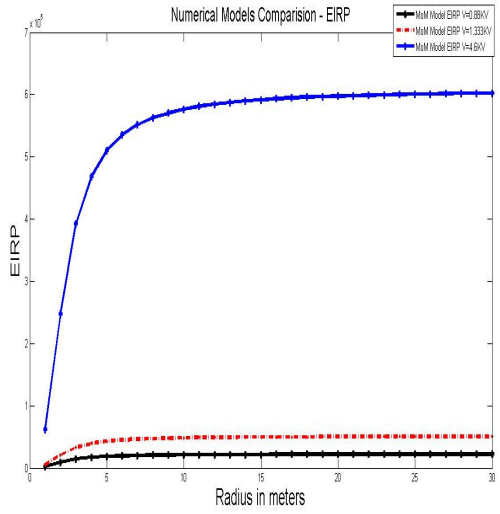


Figure 3. Values of EIRP versus Voltage feedback

In Fig. 3, there are three curves and the data are simulated to 125 meters in height. This chart shows the EIRP, (effective isotropic radiated power in Watts) and was calculated on the basis of different voltage feedback (0.88 kV, 1.333 kV and 4.6 kV) of the antenna array. It demonstrated that with a high voltage feed of 4.6 kV, the EIRP values are very high. This should give a warning to the decision makers. The two other values used in the simulation showed low values of EIRP, (less than 50 kW).

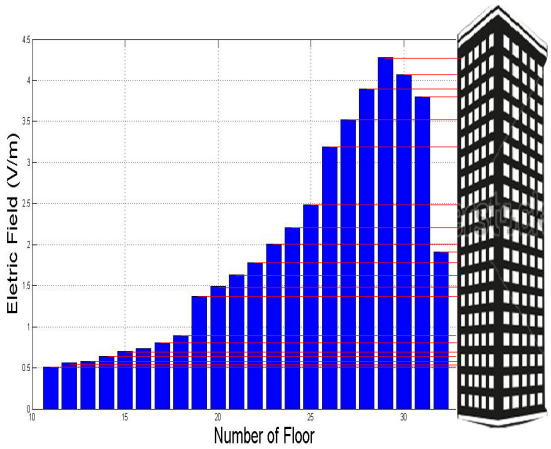


Figure 4. Electric field simulated on each floor [*In chart > Number of Floors]

The most interesting analysis is found in Fig. 4, where the electric field is simulated on each floor of a building that is higher than 125 meters and at a distance of 66 meters from the antenna array. The simulation showed the electric field increased as far as the main lobe of the antenna array and then decreased. The maximum value of the electric field was 4.28 V/m at floor number 29. This situation is similar to what happens in real life where buildings are often built near to the antenna locations or vice-versa.

V. CONCLUSION

This paper outlines the results of the effects of electromagnetic fields generated by antenna arrays. An antenna array has the advantage of providing a high gain for long distances, which means it can be used for emerging applications. One of its applications is in digital televisions and mobile devices that use smart antennas.

In this study, the effects are predicted by calculating antenna patterns using MoM, if the antenna construction is known and using the proposed model to generate the protection zone (as shown in Fig. 2). The tool also generates numerical simulations of the EIRP versus voltage feedback, as shown in Fig. 3 and finally, from a specific distance from the antenna location, it generates an electric field of any height to predict the exposure level of buildings surrounding the antenna.

The main problem in a practical application of complex computational techniques (when compared with the findings of this study) is that ray tracing, for example, requires the geometry to be specified in detail. On the othe hand, in practice, the obstacle to using even simple two-ray models, is a lack of adequate information about the antenna and the exposure level to the environment. This means that the available data about the terrain may have limited resolution. Another example is when the antenna pattern provided by the manufacturer is valid for the far-field region. Thus, it is believed that this study has made a feasible contribution to this field of studies.

The research contributions of the paper is a practical tool to calculate electric field to predict the protection zone

 ISBN: 978-1-61208-398-8

around the antenna arrays, aiding the researches without any complexity technique as ray tracing or too simplicity as two ray model, providing the researchers in this area with predicted electric fields values to aid ou to validate their works in electromagnetic compatibility and so on.

Moreover, despite the scientific uncertainty surrounding this issue and the absence of tools and practical procedures, this research has sought to continue to assist public decision making by making a calculation of the exposure levels of NIR emitted by the antennas that will be designed. The next stage is to perform signal strength measurements to confirm the validity of the developed tool.

ACKNOWLEDGMENT

The present work was done with the support of CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil. The authors also would like to thank the Laboratory for Computing and Telecommunications (LCT), Federal University of Pará (UFPA-BR), and National Institute of Science and Technology - Wireless Communication (INCT- CSF) for the support given to the research described in this paper.

REFERENCES

[1] G. International Commission on Non Ionizing Radiation Protection, “Guidelines for Limiting Exposure to Time-varying Electric, Magnetic, and Electromagnetic Fields (up to 300 GHz)”, Health Phys., vol. 74, pp . 494-522, 1998.

[2] L. Agostinho, T. B. A. Marco, and S. M. J. Antonio, “Estimating the Location of Maximum Exposure to Electromagnetic Fields Associated with a Radiocommunication Station”, Journal of Microwaves, Optoelectronics and Electromagnetic Applications, vol. 12, no. 1, June 2013, pp. 141-157.

[3] C. Guerreiro, R. Fraiha, I. Ruiz, J. Rodrigues, S. Fraiha, H. Gomes, J. Araújo, and G. Cavalcante, ”Software Tool to Aid the Definition ofC Protection Zones of Non-Ionizing Radiating in UHF Range”, 7th European Conference on Antennas and Propagation (EuCAP), vol. 1, April 2013, pp. 1003-1006.

[4] C. Guerreiro, R. Fraiha, I. Ruiz, J. Rodrigues, S. Fraiha, H. Gomes, J. Araújo, and G. Cavalcante, “Modeling electromagnetic wave propagation in the UHF range using the MoM to assess NIR exposure level”, Microwave & Optoelectronics Conference (IMOC), 2013 SBMO/IEEE MTT-S International, vol. 1, Aug. 2013, pp. 1-4.

[5] C. Chio, S. Ting, X. Zhao, T. K. Sarkar, Y. Zhang, and K. Tam, “Prediction model for radiation from base station antennas using electromagnetic simulation”, Microwave Conference Proceedings (APMC), vol. 1, Dec. 2012, pp. 1082 – 1084.

[6] A. Lala, B. Kamo, and S. Cela, "A Method of GSM Antenna Modeling for the Evaluation of the Exposed Field in the Vicinity", Network-Based Information Systems (NBiS), 2011 14th International Conference on, vol. 1, Sept. 2011, pp. 513-516.

[7] B. Thors, M. L. Strydom, B. Hansson, F. J. C. Meyer, K. Karkkainen, P. Zollman, S. Ilvonen, and C. Tornevik, “On the Estimation of SAR and Compliance Distance Related to RF Exposure From Mobile Communication Base Station Antennas”, IEEE Transactions on Electromagnetic Compatibility, vol. 50 , no. 4, Nov. 2008, pp. 837-848.

[8] M. Barbiroli, C. Carciofi, and D. Guiducci, “Assessment of Population and Occupational Exposure to Wi-Fi Systems: Measurements and Simulations”, IEEE Transactions on Electromagnetic Compatibility, vol. 53, no. 1, Feb. 2011, pp. 219-228,.

[9] D. Stratakis, A. Miaoudakis, T. Yioultsis, and T. Xenos, “Evaluation of the electromagnetic fields exposure produced by WiMAX signals”, International Conference on Telecommunications and Multimedia (TEMU), vol.1, July 2012, pp. 185-189.

[10] ANATEL, Resolution 303/2002 - regulation limiting the exposition to electrical, magnetic and electromagnetic fields in the 9 kHz to 300 GHz frequency band, in www.anatel.gov.br [retrieved: 02/2015] (in Portuguese).

[11] S. N. Makarov, Antenna and EM Modeling with Matlab, John Wiley & Sons, New York 2002.

[12] S. Makarov, “MoM antenna simulation with Matlab: RWG basis functions,” IEEE Trans. Antennas Propagation Magazine, vol. 43, no. 5, Oct. 2001, pp. 100–107.

[13] Telecommunication Standardization Sector of ITU, "Guidance on measurement and numerical prediction of electromagnetic fields for compliance with human exposure limits for telecommunication installations," International Telecommunication Union, Recommendation Rec. ITU-T K.61 (02/2008), 2008.

 ISBN: 978-1-61208-398-8

Analyzing the Optimum Switching Points for Adaptive FEC in Wireless Networks with Rayleigh Fading

Moise S. Y. Bandiri
National Institute of Telecommunications - Inatel
Santa Rita do Sapucaí, Brazil
e-mail: jumoses2000@yahoo.fr

José Marcos C. Brito
National Institute of Telecommunications - Inatel
Santa Rita do Sapucaí, Brazil
e-mail: brito@inatel.br

***Abstract*—Adaptive techniques have an important role in modern wireless communications networks, like cognitive radio networks. Adaptive modulation and adaptive Forward Error Correction (FEC) are two very important approaches used to improve the performance of the wireless networks. In an adaptive technique, a key factor to maximize the performance of the system is the optimum switching point between neighboring modulations, in an adaptive modulation system, or codes, in an adaptive FEC system. In this paper, we analyze the optimum switching points for adaptive FEC in a wireless network with Rayleigh fading. We compute the optimum switching points considering three criteria: maximum throughput, maximum packet error rate target and delay to transmit a correct packet. We show that the optimum switching points depend on several parameters, including the channel model and the selected criterion to define the switching points.**

Keywords—adaptive FEC; performance analysis; optimum switching points.

I. INTRODUCTION

A mandatory issue for the next generation of wireless networks is to improve the performance. Several technologies have been proposed to achieve this goal. One important technology is cognitive radio networks, in which the radio adjusts its parameters as a function of the radio environment in order to achieve the best performance [1].

Adaptive techniques (like adaptive modulation and adaptive error control) are fundamental to implement cognitive radio networks and also for other candidate technologies for the next-generation of the wireless networks. In particular, adaptive Forward Error Correction (FEC) schemes have been proposed to improve Quality of Service (QoS) for several technologies and techniques [2]-[7].

Adaptive FEC schemes vary the number of parity bits and, consequently, the error correction capacity of the error correcting code as a function of the quality in the wireless link.

A key point in implementing an adaptive FEC scheme is to define the optimum switching point from one error correcting code to its neighboring error correcting code. In our analysis, we consider two error correcting codes as neighboring if their error correction capacity differs by one bit. In other words, two error correcting codes are neighboring if one has an error correction capacity equal to t bits and the other one has an error correction capacity equal to $(t + 1)$ or $(t – 1)$ bits.

The optimum switching points for an adaptive FEC system have been analyzed by Brito and Bonatti in [8]. However, those analyses consider a memoryless channel, which may not be appropriate for some wireless links. In a more general case, the wireless channel has memory and the results presented in [8] are imprecise.

The goal of this paper is to analyze the optimum switching points for an adaptive FEC system considering a channel with Rayleigh fading. The criteria used to compute the optimum switching points are the same used in [8]: maximum throughput, for real time traffic, maximum Packet Error Rate (PER) target and, for non-real time traffic, the delay to transmit a correct packet.

To define the switching points it is necessary to adopt a model to calculate the PER in a Rayleigh channel. The model considered in our analysis has been presented by Sharma, Dholakia and Hassan in [9] and is summarized in Section II of this paper.

The remainder of this paper is organized as follows: in Section II, we summarize the mathematical model used to compute the PER in the wireless channel; in Section III, we analyze the optimum switching points that maximize the throughput in the wireless network; in Section IV, we consider the maximum PER target criterion; in Section V, we compute the optimum switching points considering the delay to transmit a correct packet as the QoS parameter; and finally, in Section VI, we present our conclusions.

II. MODEL TO COMPUTE THE PACKET ERROR RATE

In a Rayleigh channel, the errors tend to occur in bursts and the PER can not be computed using the Binomial distribution. Thus, it is necessary to use a model that considers the memory of the channel.

Several papers address the problem of calculating the PER in a channel with burst errors [9]-[12]. In this paper, we use the analytical model presented by Sharma, Dholakia and Hassan in [9]. In this model, the Rayleigh channel is represented by a Gilbert-Elliott Channel (GEC).

A GEC is represented by a discrete time Markov chain with two states: Good (G) and Bad (B). Figure 1 illustrates a GEC with transition probabilities α and β; each state is modeled as a Binary Symmetric Channel (BSC) with bit error probabilities p_g, in G state, and p_b in B state.

The steady state probabilities of the Markov chain illustrated in Figure 1 are given by

$$\pi_g = \frac{\alpha}{\alpha + \beta} \quad (1)$$

$$\pi_b = \frac{\beta}{\alpha + \beta} \quad (2)$$

 ISBN: 978-1-61208-398-8

where π_g and π_b are the steady state probabilities of the good and bad states, respectively.

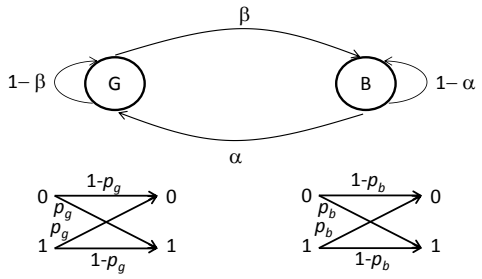


Figure 1. Gilbert-Elliott Channel.

Considering a slow fading channel, with respect to a bit interval, the probability density function of the Signal-to-Noise Ratio (SNR) is given by

$$f(\tau)=\frac{1}{\lambda}e^{\left(\frac{-\tau}{\lambda}\right)} \quad \tau>0 \tag{3}$$

where λ is the average SNR and τ is the instantaneous SNR.

The status of the channel is defined by a given threshold (ψ): the channel is considered in good state if the SNR is above the threshold and in bad state if the SNR is below the threshold.

The transition probabilities of the GEC can be calculated by [9]:

$$\alpha=\frac{f_d T\sqrt{2\pi\Gamma}}{e^{\Gamma}-1} \tag{4}$$

$$\beta=f_d T\sqrt{2\pi\Gamma} \tag{5}$$

where T is the symbol interval and f_d is the maximum Doppler speed. The parameter Γ is the ratio between the threshold (ψ) and the average SNR in the wireless channel. Following [9], in our analysis we set the parameter Γ equal to 0.1, meaning that a SNR 10 dB below the average SNR is the condition resulting in the transition from the good to the bad state.

The Bit Error Rate (BER) for each state of the channel is computed by (6) and (7) [9]:

$$p_g=\frac{\int_{\psi}^{\infty}BER(\tau)f(\tau)d\tau}{\int_{\psi}^{\infty}f(\tau)d\tau} \tag{6}$$

$$p_b=\frac{\int_{0}^{\psi}BER(\tau)f(\tau)d\tau}{\int_{0}^{\psi}f(\tau)d\tau} \tag{7}$$

where $BER(\tau)$ is the bit error rate for an Additive White Gaussian Noise (AWGN) channel with SNR equal to τ and $f(\tau)$ is the probability density function of the SNR for Rayleigh fading, given by (3).

After calculating the bit error rate by (6) and (7), we can compute the packet error rate in each state of the channel. In this paper, we are interested in analyzing the optimum switching points of an adaptive FEC system. Thus, we considered that an (n,k) error correcting block code is used in the wireless link, where k is the number of information bits and (n – k) is the number of parity bits in the block code. Defining the error correction capacity of the code as t, the PER in each state can be computed by [9]:

$$P(p_g)=1-\sum_{i=0}^{t}\binom{n}{i}(1-p_g)^{n-i}\,{p_g}^{i} \tag{8}$$

$$P(p_b)=1-\sum_{i=0}^{t}\binom{n}{i}(1-p_b)^{n-i}\,{p_b}^{i} \tag{9}$$

where p_g and p_b are the BER in good and bad states, given by (6) and (7), respectively.

Finally, the packet error rate in the Rayleigh channel is computed by

$$PER=\pi_g P(p_g)+\pi_b P(p_b) \tag{10}$$

where π_G and π_B are given by (1) and (2), respectively, and $P(p_g)$ and $P(p_B)$ are given by (8) and (9), respectively.

III. Optimum Switching Points Based on the Maximum Throughput Criterion

In this section, we analyze the switching points that maximize the throughput in the wireless link. To define the throughput we consider two factors: the overhead due to the error correcting code, expressed by the ratio between the number of information bits, k, and the total number of bits in a packet, n; and the percentage of packets received without error or containing only correctable errors, computed by 1 – PER. Thus, only packets received without error, after the error-correcting code action, are considered when computing the throughput (some authors refer to this as 'goodput'). As our definition of throughput does not consider the bandwidth of the channel and the modulation used in the wireless link, we refer to this throughput as a normalized throughput. Thus, the normalized throughput in the wireless link is given by:

$$T=\frac{k}{n}\cdot(1-PER) \tag{11}$$

The parameter n depends on the current code used in the adaptive FEC system and PER is the packet error rate for this code, given by (10).

To compute the parameter n as a function of the error correction capacity, t, of the current code, it is necessary to define a particular family of codes or to use some known bound. In this paper, we opted to use the bound for the Bose, Chaudhuri and Hocquenghem (BCH) code. The BCH codes form a large class of powerful random error-correcting cyclic codes. The BCH bound can be summarized as: for any positive integers m ($m \geq 3$) and t ($t < 2^{m-1}$) exists a (n,k) binary t-error correcting BCH code with the following parameters: [13]

 ISBN: 978-1-61208-398-8

$$n = 2^m - 1$$
$$(n-k) \le mt \quad (12)$$
$$d_{\min} \ge 2t + 1$$

where d_{min} is the minimum distance of the code.

In order to compare our results with those presented in [8], we set k = 424 bits when computing the PER. For this value of k, a BCH code with suitable natural length or suitable number of information digits may not be obtained. However, by subtracting a proper number of bits from the natural code, a shortened BCH code can be implemented [13].

Thus, using the bounds given by (12) and considering k = 424 bits, we can compute the number of parity bits (and thus the value of n) as a function of the error correction capacity of the code, t, as:

$$(n-k) = 9t \text{ if } 1 \le t \le 9$$
$$(n-k) = 10t \text{ if } 10 \le t \le 59 \quad (13)$$

We are interested in computing the throughput as a function of the SNR in the channel. For this we follow the steps:

1. Define a given modulation (reference modulation) to the wireless channel. We use Quadrature Amplitude Modulation (QAM) with 256 points in its constellation (256-QAM modulation)
2. Compute the BER as a function of the Symbol Energy to Noise Density Ratio (E_S/N_0) in an AWGN channel, using classical formulas presented in the literature, like in [14].
3. Using (6) and (7), compute the BER in the good and bad states.
4. Define the error correction capacity of the current code and, using (13), the number of parity bits in the code and after the total number of bits, n.
5. Using (8), (9) and (10), compute the PER in the Rayleigh channel.
6. Finally, compute the normalized throughput for the current code using (11).

Figure 2 shows the normalized throughput as a function of E_S/N_0 considering codes with different error correction capacity, a 256-QAM in the wireless link and k = 424 bits. The optimum switching point between two neighboring codes is obtained by the cross point of the corresponding throughput curves.

Table I summarizes the optimum switching points presented in Figure 2 and, to permit comparisons, the optimum switching points previously published in [8].

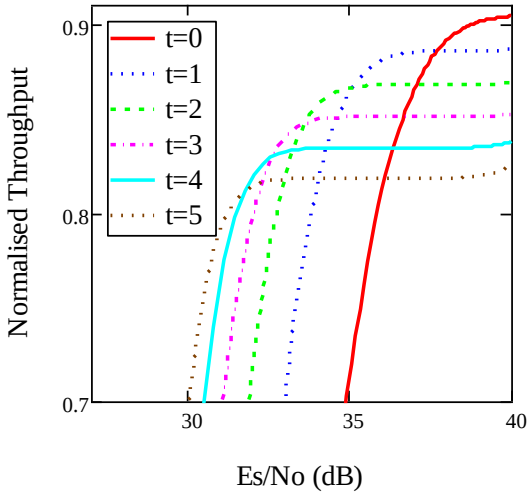


Figure 2. Normalized throughput as a function of *Es/No* for k = 424 bits and 0 ≤ t ≤ 5.

TABLE I. OPTIMUM SWITCHING POINTS FOR MAXIMUM THROUGHPUT CRITERION, K = 424 BITS.

Switching		E_s/N_0 (dB) obtained from Fig. 2	E_s/N_0 (dB) obtained from [8]
From	to		
t = 0	t = 1	37.7	30.7
t = 1	t = 2	35.1	29.0
t = 2	t = 3	33.6	28.1
t = 3	t = 4	32.6	27.5
t = 4	t = 5	31.8	27.0

Based on the results presented on Table I, we can conclude that the optimum switching points depend on the channel model.

Another interesting point is to investigate the influence of the packet length, k, in the optimum switching points. For this, we compute again the normalized throughput considering now k = 848 bits (note that, with this new value of k, the Equation (13) needs to be modified based on the bounds presented on (12)). Figure 3 shows the new results obtained and Table II compares the optimum switching points for k = 424 bits and k = 848 bits.

 ISBN: 978-1-61208-398-8

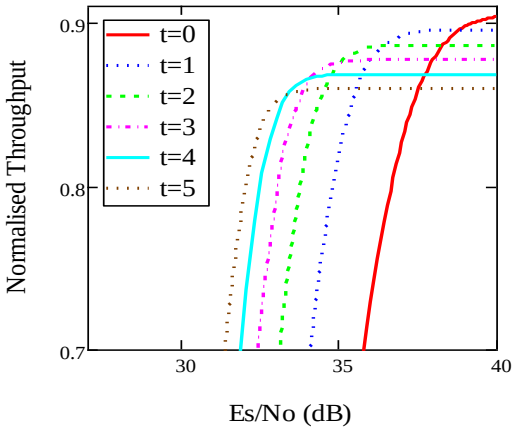


Figure 3. Normalized throughput as a function of *Es*/*No* for k = 848 bits and 0 ≤ t ≤ 5.

TABLE II. OPTIMUM SWITCHING POINTS FOR MAXIMUM THROUGHPUT CRITERION, K = 848 BITS.

Switching		E_s/N_0 (dB) k = 848 bits	E_s/N_0 (dB) k = 424 bits
from	to		
t = 0	t = 1	38.9	37.7
t = 1	t = 2	36.4	35.1
t = 2	t = 3	35.1	33.6
t = 3	t = 4	34.1	32.6
t = 4	t = 5	33.4	31.8

Based on the results summarized in Table II, we can conclude that the optimum switching points depend on the number of information bits in the packet, with the E_S/N_0 in the switching points increasing when k increases.

It is important to observe that the PER in the optimum switching points can be unacceptable for some applications. For example, considering k = 848 bits, the PER is about 0.11 in the switching points. This result leads us to the next criterion to define the optimum switching points: the maximum PER target.

IV. OPTIMUM SWITCHING POINTS BASED ON THE MAXIMUM PER TARGET

The criterion presented in the last section maximizes the throughput but does not consider any restriction about PER. If the application needs to maintain the PER below a given threshold, the maximum PER target is a better criterion. In this criterion we compute the PER for each error correcting code, as a function of the E_S/N_0. The optimum switching point in this case is the cross point between the PER curve and the PER threshold. To illustrate this approach, Figure 4 shows the switching points considering the maximum PER target equal to 0.15 (actually, this value is too high for most of the applications and has just been chosen to illustrate the concept). To plot Figure 4 we set k = 424 bits and we consider 256-QAM modulation. Table III summarizes the switching points showed in Figure 4.

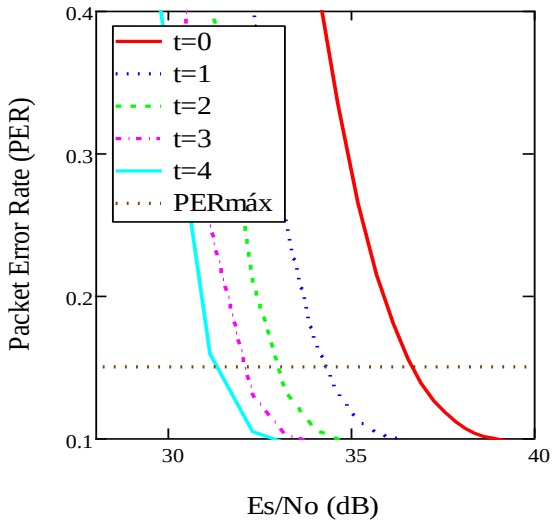


Figure 4. Switching points considering maximum PER criterion with target equal to 0.15 for 0 ≤ t ≤ 5 .

TABLE III. OPTIMUM SWITCHING POINTS FOR MAXIMUM PER TARGET.

Switching		E_s/N_0 (dB)
from	to	
t = 0	t = 1	36.7
t = 1	t = 2	34.3
t = 2	t = 3	32.9
t = 3	t = 4	32.0
t = 4	t = 5	31.3

Based on the results presented in Table III and comparing these results with those presented in Section II and those presented on reference [8], we can conclude that: the optimum switching points depend on the selected criterion; again, the optimum switching points depend on the channel model.

V. OPTIMUM SWITCHING POINTS BASED ON THE DELAY

In this section, we consider the mean delay to transmit a correct packet as the criterion to define the optimum switching points. This criterion is appropriate for non-real time applications where we can retransmit a packet in order to guarantee that all packets delivered to the receiver are correct.

To compute the delay it is necessary to define a reference scenario in terms of multiple access and the strategy to retransmit packets. In order to permit comparisons, we use here the same reference scenario used in [8]:

- A TDMA (Time Division Multiple Access) system with X time slots in a frame, whit n bits being transmitted in each time slot.
- Each packet is transmitted over Z frames (one packet needs Z time slots to be transmitted).
- Each packet is retransmitted until being correctly received (we consider that the number of retransmissions in unlimited).

 ISBN: 978-1-61208-398-8

- A slow fading channel, meaning that the duration of the fades is greater than de duration of the packet transmission.
- Each time slot is protected by an error correcting code, but the retransmissions occur at the packet level and not at the time slot level. In other words, the receiver requests retransmission of the whole packet.

With these assumptions, the mean delay to transmit a correct packet is given by [8]

$$E(T) = [(Z-1)X+1]\frac{n}{R}(1-PER) \quad (14)$$

where R is the transmission rate in the wireless link and the parameter PER is calculated at the packet level. As the FEC code is applied in each slot, we need to modify (8) and (9) to compute the PER associated with each state of the GMC. The new equations are:

$$P(p_g) = 1-\left(\sum_{i=0}^{t}\binom{n}{i}(1-p_g)^{n-i}\,p_g{}^{i}\right)^{Z} \quad (15)$$

$$P(p_b) = 1-\left(\sum_{i=0}^{t}\binom{n}{i}(1-p_b)^{n-i}\,p_b{}^{i}\right)^{Z} \quad (16)$$

With these new equations we can apply (10) to compute the PER and then we can calculate the delay using (14).

As we are only interested in comparing the delay for different error correcting codes and as the parameters Z, X and R are independent of the FEC code, we can define a normalized delay as:

$$E(T) = n(1-PER) \quad (17)$$

Figure 5 shows the delay for $0 \le t \le 5$, considering k = 424 bits and Z = 5. For each code, n is computed using (13) and PER is computed using (15), (16) and (10). Again, the optimum switching point between any two neighboring codes is obtained by the cross point of the corresponding curves illustrated in the figure.

Table IV summarizes the optimum switching points obtained from Figure 5. Comparing these results with those presented on previously sections, we can conclude that the optimum switching points depend on the selected criterion. Again, comparing these results with those presented in [8], we conclude that the optimum switching points depend on the channel model.

TABLE IV. OPTIMUM SWITCHING POINTS FOR DELAY CRITERION.

Switching		E_s/N_0 (dB)
from	to	
t = 0	t = 1	39.0
t = 1	t = 2	36.2
t = 2	t = 3	34.6
t = 3	t = 4	33.5
t = 4	t = 5	32.7

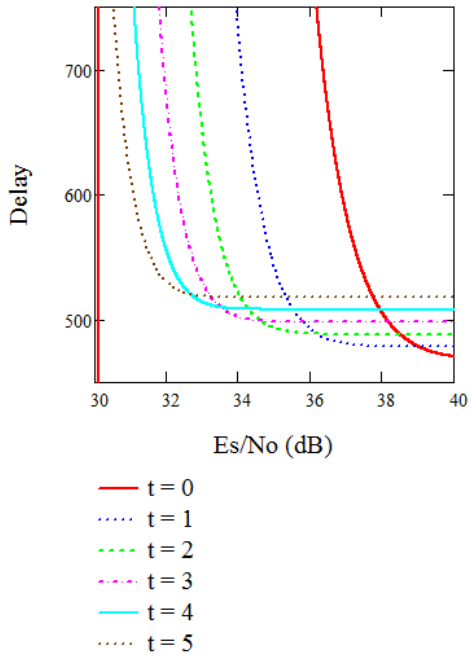


Figure 5. Switching points considering delay criterion for $0 \le t \le 5$.

VI. CONCLUSIONS

The switching point is a key factor to maximize the performance of systems using adaptive FEC. The optimum switching points for this kind of system have been analyzed in [8] considering three criteria: maximum throughput, maximum PER target and delay to transmit a correct packet. However, analyses performed in [8] consider an AWGN channel or that an interleaving process is used to randomize the burst errors in the wireless channel.

In this paper, we analyze the optimum switching points considering a wireless link modeled as a Rayleigh channel. We use the same three criteria proposed in [8]. We concluded that the optimum switching points depend on: the channel model, the selected criterion and the packet length.

In future works, we will extend the results presented here considering an adaptive hybrid system, in which we combine adaptive modulation and adaptive FEC in order to improve the performance of the system. Also, we will consider different fading models, like Rician model and Nakagami model.

REFERENCES

[1] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty, “NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey”, Computer Networks 50 (2006), pp. 217-2159.

[2] J. Yang, A. K. Khandani, and N. Tin, “Adaptive Modulation and Coding in 3G Wireless Systems”, Proceedings on 56th Vehicular Technology Conference, VTC 2002-Fall, pp. 544-548.

[3] S. H. O. Salih and M. M. A. Suliman, “Implementation of Adaptive Modulation and Coding Technique using”, International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011, pp. 137 – 139.

[4] Y. Fakhri, B. Nsiri, D. Aboutajdine, and J. Vidal, "Adaptive Throughput Optimization in Downlink Wireless OFDM System," Int'l J. of Communications, Network and System Sciences, Vol. 1 No. 1, 2008, pp. 10-15.

 ISBN: 978-1-61208-398-8

[5] M. Taki and F. Lahouti, "Spectral Efficiency Optimization for an Interfering Cognitive Radio with Adaptive Modulation and Coding", IEEE J. Select. Areas Commun, 2009, pp.1-6.

[6] M.R.Ebenezar jebarani and T.Jayanthy, "An Analysis of Various Parameters in Wireless Sensor Networks Using Adptive FEC Technique", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.1, No.3, September 2010, pp.33-43.

[7] Y.Baguda et al., "Adaptive FEC Error Control Scheme for Wireless video Transmission", In: Advanced Communication Technology (ICACT), 2010 The 12th International Conference on. IEEE, 2010, pp. 565-569.

[8] J. M. C. Brito and I. S. Bonatti, "Analysing the switching points in wireless ATM Networks that use adaptive FEC Schemes", Proceedings the second International Symposium on Communications and Information Technology, Pattaya, Thailand, 23-25 October 2002, pp. 305-308.

[9] G. Sharma, A. Dholakia, and H. Hassan, "Simulation of error trapping decoders on a fading channel", IEEE Vehicular Technology Conference, GS, Atlanta, USA, VOL 2, 28 Apr. - 1 May 1996, pp. 1361-1365.

[10] R. Khalili and K. Salamatian, "A new analytic approach to evaluation of packet error rate in wireless networks", Proceeding of the 3rd Communication Networks and Services Research Conference, 2005, pp. 333–338.

[11] Y. Xi, A. Burr , J. Wei, and D. Grace, "A general upper bound to evaluate packet error rate over quasi-static fading channels", IEEE Transactions on Wireless Communications, V.10, No 5, May 2011, pp. 1373-1375.

[12] C. Jiao, L. Schwiebert, and B. Xu; "On modeling the packet error statistics in bursty channels", Proceedings on 27th Local Computer Networks - LCN, November 2002, pp. 534-541.

[13] S. Lin and D. J. Costello; Error Control Coding: Fundamentals and Applications. Prentice Hall, 1983.

[14] B. Sklar, Digital Communications – Fundamentals and Applications, 2nd edition, Prentice Hall, 2001.

 ISBN: 978-1-61208-398-8

Multicast Receiver Access Control in the Automatic Multicast Tunneling (AMT) Environment

Veera Nagasiva Tejeswi Malla
J. William Atwood

Department of Computer Science and Software Engineering
Concordia University, Montreal, Quebec, Canada
Email: tejam.vns@gmail.com, william.atwood@concordia.ca

***Abstract*—The Automatic Multicast Tunneling protocol extends the range of multicast data distribution from a multicast-enabled network region to a network region that supports only unicast routing. Previous work has shown how to achieve access control in network regions that fully support multicast routing. In this paper, we show how to achieve the access control in the extended (unicast-only) network region, without modifying the original interactions of the access control. We also formally validate the security of our solution using the Automated Validation of Internet Security protocols and Applications (AVISPA) tools.**

***Index Terms*—Automatic Multicast Tunneling; Access Control; Unicast Network; Multicast Network.**

I. INTRODUCTION

Some applications require data to be delivered from a sender to multiple receivers. Examples of such applications include audio and video broadcasts, real-time delivery of stock quotes, and teleconferencing. A service where data are delivered from a sender to multiple receivers is called multipoint communication or Multicast. It provides an efficient way to support high bandwidth, one-to-many applications on a network. One major problem in IP multicast is that even hosts without any permissions are able to join multicast groups, i.e., there is no mechanism to prevent unauthorized users from accessing a multicast network. Consequently it became impossible for service providers to justify billing for multicast data usage.

To overcome the problem of revenue generation, *Participant Access Control* (PAC) was introduced in [1]. PAC includes *Sender Access Control* (SAC) [2] and *Receiver Access Control* (RAC) [3], [4]. RAC is a scalable, distributed and secure architecture, where authorized end users can be authenticated before delivering any data. Although PAC provides access control for IP multicast, it is limited to native multicast environments.

To overcome the requirement to support native multicast routing, a solution was proposed by the Internet Engineering Task Force (IETF) called Automatic Multicast Tunneling (AMT). Without requiring any manual configuration, AMT allows a device in a network region supporting only unicast routing to receive multicast traffic from the native multicast infrastructure. The goal of AMT is to provide a migration path from no multicast support to full multicast support, and thus foster the deployment of native IP multicast. An Internet Service Provider can offer AMT-based service until such time as the number of multicast-capable customers justifies the expenditure for multicast-capable routers. Although AMT provides a simple-to-implement way to improve multicast availability, it provides no RAC for multicast groups.

In this paper, we have proposed a design architecture that provides RAC in AMT. We have also formally validated the security features of our model using the Automated Validation of Internet Security protocols and Applications (AVISPA) tool [5].

The rest of the paper is organized as follows: Section II gives background information on the PAC architecture, the Internet Group Management Protocol (IGMP), the Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol, the Extensible Authentication Protocol (EAP), the Protocol for Carrying Authentication for Network Access (PANA), the Secure IGMP (SIGMP) protocol, the Group Security Association Management (GSAM) protocol, and AMT. Section III provides the problem definition. Section IV defines our proposed solution. Section V discusses some alternate approaches. Section VI shows how we have modeled our solution using the AVISPA formal modeling tool. Section VII concludes our paper.

II. BACKGROUND

In this section, we first present the PAC architecture for native IP multicast that was developed within our group. This is followed by a brief description of the related protocols.

A. PAC Architecture

The architecture shown in Figure 1 was proposed in [6]. A number of parties that participate in a multicast session, either before the session or during it, have been identified. The *Content Provider* offers the product to be delivered to the multicast group. The *End User* (EU) receives the content. The *Network Service Provider* (NSP) delivers the data, making use of *Access Routers* (ARs), *Core Routers* (CRs), one or more instances of an *Authentication, Authorization and Accounting*

 ISBN: 978-1-61208-398-8

Server (AAAS), and a *Network Access Server* (NAS) associated with each AR. We will assume that the ability of the EU to pay for services will be certified by a *Financial Institution* (FI). The *Group Owner* (GO) is responsible for the creation and overall activities of the group. PAC can be further divided into SAC and RAC.

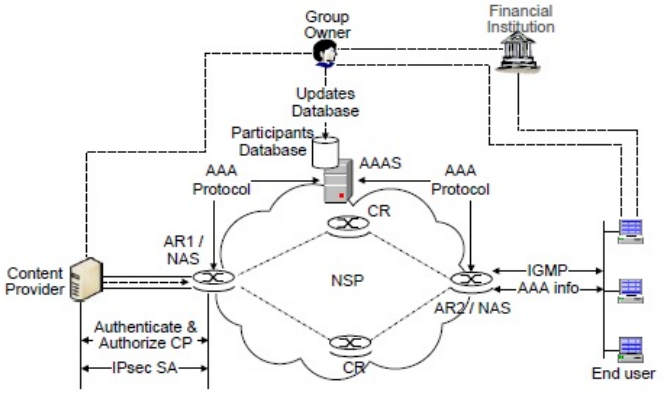


Fig. 1. Access Control Architecture for Multicast Participants.

SAC will be deployed at the interface between the CP and the network, where AR1 will authenticate and authorize the CP after an interaction with the AAAS. On successful authentication and authorization, an Internet Protocol Security (IPsec) *Security Association* (SA) [7] will be established between the CP and AR1 to cryptographically authenticate each data packet before forwarding it to the multicast distribution tree [2]. As AMT is targeted for extending the options for receivers, we assume that the Content Provider is in a multicast-capable region, and will not discuss SAC any further.

RAC will be implemented at the interface between the network and the EU's device. AR2 will receive and process the network level join (IGMP) messages (see Section II-B) and the messages carrying *Authentication, Authorization and Accounting* (AAA) information (see Section II-C and Section II-D). It will also act as a NAS by communicating with the AAAS. It is assumed that the Group Owner has supplied the user authentication information or AAA information to the AAAS when the EU purchased the service. Hence, each EU will be authenticated and authorized by the one-hop AR before allowing him/her to join a secured group [3], [4]. Several IPsec SAs will be established to cryptographically authenticate the Secure IGMP messages (see Section II-E) exchanged between the EU device and the network [8].

B. IGMP and PIM-SM

IGMP [9] has been standardized by the IETF for IPv4 systems (*host* or *router*) to inform the neighboring router(s) about the multicast group memberships of these systems. IGMP performs three main operations:

- A host sends a *join* message (through a Membership Report Message) when it wants to join a multicast group or some specific sources of a group.
- A host sends a *leave* message (through a Membership Report Message) when it wants to unsubscribe from a multicast group
- A router periodically checks (through a Membership Query Message) which multicast groups are of interest to the hosts that are directly connected to that router.

While IGMP is the protocol used between an EU host and its AR, a multicast routing protocol (typically PIM-SM [10]) is used to build the data distribution tree among the CRs and the ARs. An IGMP join message will cause the grafting of one or more new edges (if there are no existing clients on the same AR for that group) and an IGMP leave message will cause the data distribution tree to be pruned if this is the last client on the AR.

C. EAP

To achieve AAA functions, a AAA protocol (e.g., Diameter [11]) is used between a NAS and its associated AAAS. The specific aspects of (EU) authentication and authorization are typically delegated to EAP [12], which is a versatile framework that facilitates the use of multiple authentication methods, such as pre-shared secret, one time password, public key authentication, etc. Although EAP was originally intended to be used to control access to a network as a whole, it is also useful for managing access at the application layer. In particular, EAP procedures can be adapted for use in multicast-based applications, to authenticate the users, to authorize them, and to account for their group-level activity [6]. EAP does not run directly over the IP layer. The mechanism for carrying the EAP packets will be discussed in Section II-D.

The EAP framework supports multiple authentication mechanisms called *methods*. EAP runs between an *Authenticator* (on the AR) and a *Peer* (on the EU host). The Authenticator normally acts as a pass-through to a back-end *Authentication Server* (AS), which will be co-located with the AAAS. The EAP packets that arrive at the NAS are sent to the AAAS by encapsulating them inside AAA packets, and the NAS will decapsulate the AAA packets received from the AAAS and forward the EAP packets to the Peer on the EU host.

A justification for using the method EAP-FAST in our application may be found in [13].

D. PANA

The IETF has standardized Protocol for carrying Authentication for Network Access (PANA) [14], a protocol that carries EAP authentication methods (encapsulated inside EAP packets) between a client node (EU host) and a server in the access network.

The PANA framework [15], comprised of four functional entities, is shown in Figure 2. The *PANA Client* (PaC) residing on a requesting node (e.g., an EU host) interacts with the *PANA Authentication Agent* (PAA) in the authentication process using the PANA protocol [14]. The server implementation of PANA is the PAA, which consults an *Authentication Server*

 ISBN: 978-1-61208-398-8

(AS) for authentication and authorization of a PaC. If the PAA is separate from the AS, a AAA protocol (e.g., Diameter) will be used for their communication. The PAA resides on a node that is typically a NAS in the access network. The AS is a conventional back-end AAAS that terminates the EAP and the EAP methods. The *PANA Enforcement Point*

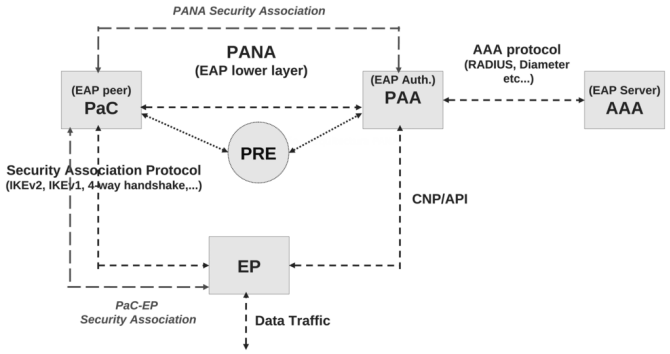


Fig. 2. PANA Framework.

(EP) allows (blocks) data traffic of authorized (unauthorized) clients. When the PAA and EP reside on the same node, they use an API for communication, otherwise, a protocol (e.g., SNMP) is required. A secure association protocol (e.g., IKEv2 [16]) is required to run between the PaC and the EP to establish an IPsec [7] Security Association (SA) [17], which can provide integrity protection, data origin authentication, replay protection and optional confidentiality protection.

When EAP-FAST or an equivalently-capable EAP method is used, a shared key becomes available to the PAA and the PaC. To establish an indirect coupling between the PANA/EAP-based authentication and IGMP join/leave operations, the shared key (or a key derived from that shared key) established during the PANA session can be used to protect IGMP messages (following the security guidelines of the IGMPv3 [9] specification).

E. SIGMP

The Secure Internet Group Management Protocol (SIGMP) [8] is an extension of IGMP, which runs among the EU hosts and the ARs. It distinguishes two types of multicast groups: *open* groups, which are identical to mulitcast groups with IGMP, and *secure* groups, for which the exchanges are protected. As for IGMP, in SIGMP the EU host implements the host portion of SIGMP while the AR implements the router portion of SIGMP. SIGMP queries and reports are each divided into two categories, Open Group Query (OGQ), Secure Group Query (SGQ), Open Group Report (OGR), Secure Group Report (SGR). OGQ and OGR are for open groups and SGQ and SGR are for secure groups. In SIGMP, queries and reports for open groups are delivered without any protection (i.e., exactly as they would be for IGMP), but for secure groups they are protected by IPsec Group Security Associations (GSAs). Two different GSA instances are used: GSA_q and GSA_r. GSA_q is used to protect the SGQ messages and GSA_r is used to protect the SGR messages.

F. GSAM

The Group Security Association Management (GSAM) protocol is used to manage the GSAs used in SIGMP (similar to IKEv2 in unicast). The network entities in GSAM are the same as those in SIGMP, including ARs and EU hosts. In IGMP (and SIGMP), if there are multiple routers on a network segment, one of them will be elected as the *Querier* (Q), and the remaining routers are called *Non-Queriers* (NQ). In GSAM, an AR (specifically, the Querier) plays the role of *Group Controller / Key Server* (GCKS). It accepts registrations from NQs and EUs that have been authorized at the application level and grants them group membership in the secure multicast groups that the EUs are authorized to join. The members of this set of EUs are called *Group Members* (GMs). The AR/Q creates and updates GSA_r and GSA_q for a secure group and distributes them to GMs in the secure group using secure tunnels. The Q, the NQs (if any), and the GMs will update their local IPsec databases (Security Association Database (SAD) and Group Security Policy Database (GSPD)) [7] according to the parameters of GSA_q and GSA_r to protect the SIGMP packets (for more details about SIGMP, GSAM and their interactions, see [8]).

G. RAC System Operation

The operation of RAC can be viewed at two levels: the application level and the network level.

At the application level, an EU will negotiate with the GO to obtain permission to access a particular product (e.g., a video stream). After consulting with the FI to determine the ability of the EU to pay, the GO will issue a “ticket” to the EU, which describes how and when to connect to the stream representing the product (i.e., it gives the network-level group address), and certifies the EU’s right to access the group. The *form* of this ticket is simultaneously (or has been previously) provided to the NSP, to permit rapid validation.

The ticket is presented to the NSP by the EU, using EAP [12]. PANA [15] is used to carry the EAP exchanges between the EU and the AR. In PANA, the PANA client (PaC) will be on the EU host. On the NSP side of the network segment, there are two PANA-related functions: the PAA and one or more EPs. The EPs are ultimately responsible for enforcing the restrictions in a network-level join. If an EU is not authorized, then a network-level join request from that EU’s host will be blocked, i.e., it will not result in a join operation in the multicast routing protocol. In the simple case (only one AR for the network segment), the PAA and the EP will be co-located with the AR. In more complex cases (more than one AR for a specific network segment), one AR will have both PAA and EP functions, and the rest will have only the EP function. An appropriate secure protocol is used to carry information from the PAA to the EPs. A AAA protocol (e.g., Diameter [11]) is used to carry the EAP exchanges between the AR and the

 ISBN: 978-1-61208-398-8

AAAS, where the ticket is validated. From the perspective of Diameter, the AR acts as a NAS, i.e., as a Diameter client.

Note that the election of Q for a network segment is independent of the designation of the PAA for that segment, so there is no pre-defined relationship among the PAA, the Q, the NQ (if any), and the NAS, although we do assume that the PAA resides on an AR that can act as a NAS.

As a result of the EAP exchanges, a PANA Master Session Key (MSK) becomes known to the PAA and the PaC. As defined in [18], the PAA must combine the MSK with EP-specific information to produce the PaC-EP Master Key (PEMK), which is then forwarded (securely) to the EP. As shown in [4], the EP must, in turn, combine this PEMK with group-specific information to produce the Multicast Session-Specific Key (MSSK), which will be used to protect the PaC-EP communications, and the (group-specific) network-level exchanges between an EU's host and its EP [8].

Note that since the MSK is specific to the multicast session, presentation of a ticket for a different session will result in the establishment of a new PANA session, a new MSK, and derivation of a new PEMK and a new MSSK.

The network-level join operation is requested through our secure extension to IGMP (see Section II-E). SIGMP is compatible with all existing versions of IGMP, and utilizes exactly the same packet formats.

The necessary security features are achieved using IPsec [7] and the Multicast Extensions to IPsec [19]. As noted above, the security parameters are derived from the MSK. The key management and coordination functions needed by SIGMP are provided by GSAM (see Section II-F).

H. Automatic Multicast Tunneling

Automatic Multicast Tunneling (AMT) [20] allows multicast communication to take place from one or more sources that have native multicast connectivity to hosts, sites or applications that do not have native multicast access, i.e., to request and receive Source Specific Multicast (SSM) or Any Source Multicast (ASM) traffic from within a network that does provide multicast connectivity. Without requiring any manual configuration, AMT allows the hosts to receive multicast traffic from the native multicast infrastructure. AMT operates with a pseudo interface, where UDP-based encapsulation is done to overcome problems of multicast connectivity.

We assume that the multicast-enabled ISP provides the *AMT Relay* service. As shown in Figure 3, the hosts connected to the unicast-only network are acting as *AMT Gateways*.

1) When host wants to join a multicast group, it sends a membership report to the Gateway thinking that it is an IGMP router (Querier).
2) Before forwarding the received report, the Gateway will send a Request message to the Relay to solicit a General Query response. The Relay responds by sending a Membership Query message back to the gateway. The Membership Query message carries an encapsulated general query that is processed by the IGMP or MLD protocol implementation on the Gateway to produce a membership/listener report. Each time the Gateway receives a Membership Query message, it starts a timer whose expiration will trigger the start of a new Request. This timer-driven sequence is used to mimic the transmission of a periodic General Query by an IGMP/MLD router. This query cycle may continue indefinitely, once started by sending the initial Request message.
3) After receiving the general query from the Relay, the Gateway will send the membership report encapsulated to the Relay. Each report is encapsulated and sent to the Relay after the Gateway has successfully established communication with the Relay via a Request and Membership Query message exchange.
4) The AMT Relay will decapsulate the IGMP messages and trigger an upstream PIM join towards the source.
5) Finally the requested multicast data are transferred from the multicast source to host through the Relay and the Gateway.

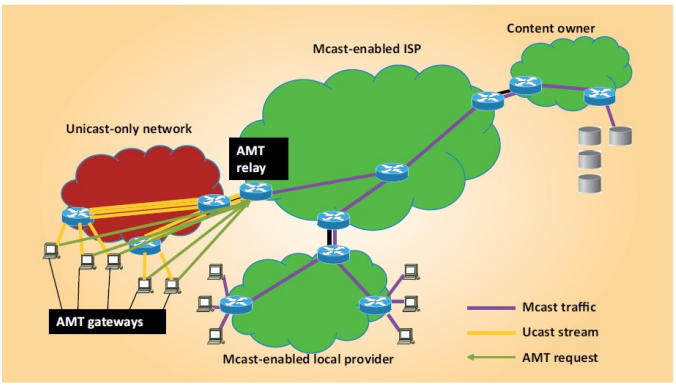


Fig. 3. AMT Architecture.

AS noted in Section I, AMT is intended as an interim measure [20]. Its purpose is to provide a (relatively) low-cost mechanism that will allow the set of multicast subscribers to grow gracefully, until the point is reached where full multicast routing support can be justified. As such, considerations of efficiency and scalability are not key issues in the design of AMT. (Any tunneling-based solution will always be less efficient than a solution that does not involve tunnels.) Similarly, scalability is not a key issue, because once the subscriber base becomes large enough for scalability to be an issue, the justification for full multicast routing support will be there.

III. Problem Definition

As previously noted, native IP multicast offers scalable point-to-multipoint delivery, but no access control. AMT extends IP multicast service to a unicast-only region, but offers no access control. The PAC environment offers access control, but is limited to the native IP multicast environment. So, our goal is to combine both, i.e., in addition to the current features of AMT, we must add RAC features. This must be

 ISBN: 978-1-61208-398-8

done without changing the interactions that are expected by the EU or the network components that reside in the native IP multicast region.

IV. Proposed Solution

As noted in Section II-H, the AMT Relay and the AMT Gateway implement the host and router portions of the IGMP interaction, respectively. Our design is based on extending the functionality of the AMT Relay and the AMT Gateway so that the EAP, PANA, SIGMP, and GSAM interactions in the AMT environment are identical to what they would be in the native IP multicast environment. Here, in this section, we explain how the RAC framework is accommodated into the AMT environment to achieve Receiver Access Control. Figure 4 shows the AMT architecture with "Receiver Access Control". The whole design is based on the fact that all messages and data between the EU host and the AR must pass through the AMT Tunnel, i.e., between the Gateway and the Relay.

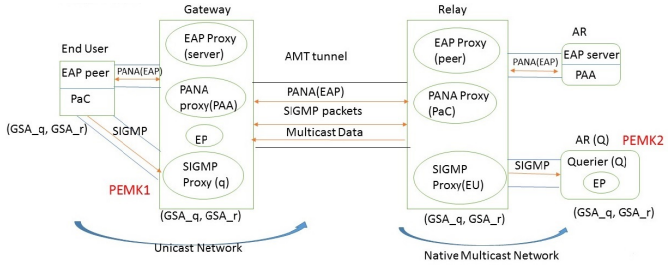


Fig. 4. Receiver Access Control in AMT.

A. EAP, PANA, SIGMP, GSAM Proxies

The need for all exchanges between the EU host (in the unicast-only region) and the AR (in the multicast-enabled region) to flow through the AMT tunnel implies that proxies must be established for all messages associated with the RAC functionality. We introduce a proxy in the Gateway for each message type; each proxy responds as if it were the AR. We introduce a corresponding proxy in the Relay, each acting as if it were on an EU host. The necessary interactions among the EAP proxy, the PANA proxy, the SIGMP proxy, and the GSAM proxy are simplified because they are all co-located in the Gateway.

Triggered by the need to send the first EAP message, the (real) PaC discovers its proxy PAA in the Gateway using the normal mechanism for PAA discovery as defined in [21], and establishes a secure relationship with it. The proxy PaC in the Relay discovers the real PAA and establishes a secure relationship with it. In effect, the Gateway and the Relay act as a "friendly" Man-in-the-Middle.

After authentication, the EAP method exports a Master Session Key (MSK) to the PaC and the PAA. As a result the EAP proxy parts in the Gateway and the Relay will know (or be able to construct) the MSK for protecting SIGMP messages.

SIGMP on the EU host interacts with the proxy SIGMP on the Gateway. It will use a GSA derived from the MSK in the same way as it would if it were in a native IP multicast environment. Similarly, the proxy SIGMP on the Relay interacts with the Querier in the native IP multicast region.

GSAM on the EU host uses the keys derived from the MSK and the proxy GSAM identity to form the necessary GSAs to protect the SIGMP exchanges between the EU host and the Gateway, and proxy GSAM on the Relay uses the keys derived from the MSK and the Querier identity to form the necessary GSAs to protect the SIGMP exchanges between the Relay and the Querier. Although the MSK has the same value, the EP identity is different in the two cases, so the derived keys will differ.

As a result, communication between the EU host and the multicast-network-based components will take place on three segments: EU host to Gateway, Gateway to Relay, and Relay to AR.

B. RAC in AMT System Operation

From the perspective of the EU, the operations proceed exactly as they would in a native IP multicast environment. However, our desired proof of security must take into consideration the existence of additional participants in these exchanges. When a SIGMP message is to be sent by the EU host for the first time, GSAM is invoked to negotiate the cryptographic parameters. This negotiation will be between the EU host and the Gateway. It will in turn trigger (through the AMT tunnel) another negotiation between the Relay and the AR in the multicast-enabled region. Further details may be seen in [22].

The RAC can be viewed at two levels: the application level and the network level.

1) Access Control at the Application Level: A PANA session consists of five phases [14]. We explain below how the PANA messages are exchanged during these phases in AMT using the PANA proxy and the EAP proxy.

1) Handshake Phase: The PaC, on receiving a request from the upper layer to join a multicast group, initiates a PANA session by sending a PANA Client Initiation (PCI) message to the Gateway thinking it is the PAA. The Gateway finds it as a PANA packet and forwards it to the Relay. The Relay, having a PANA proxy acting as a PaC, forwards the packet to the actual PAA. The response goes back from actual PAA to PaC through the Relay and the Gateway.
2) Authentication and authorization phase: After the handshake phase, EAP packets carried by PANA will be exchanged between the PaC and the PAA. For better understanding, we took an example of EAP-FAST method [23], an efficient EAP method. This method has two phases, in which phase 1 is responsible for TLS handshake resulting in a secure tunnel between peer

 ISBN: 978-1-61208-398-8

and server. As explained, the EAP proxy acting as an EAP server is in the Gateway and the EAP peer is in the EU. The secure tunnel is formed between the EU and the Gateway (say STunnel1), resulting in a fresh secret key between them. The same secure tunnel with another key is formed between the Relay and the PAA (say STunnel2) during phase 1. In phase 2, EAP method payloads carrying user credentials in PANA packets are transferred to the Gateway through STunnel1 and the Gateway, who shares the secret key with the EU during phase1, will decrypt and forward them to the Relay through the AMT Tunnel (assuming AMT tunnel is secured). Finally the Relay protects the payloads with keys obtained during formation of STunnel2 and forwards the EAP message to the PAA. The PAA verifies those credentials and authenticates EU and sends the results back.

After a successful authentication, the PaC and PAA derive a Master Session Key (MSK). As the Gateway and the Relay are part of PANA exchanges and acting as a friendly Man-in-the-Middle, they can compute the MSK as well. On receiving the MSK the PAA transfers MSK to EPQ (Enforcement point in Querier) using IPsec, with a key calculated in the normal way for two IPsec peers [24].

3) Access Phase: PaC and EPG (Enforcement point in Gateway), Relay and EPQ with acquired pre-shared key (MSK) during authentication phase calculate the secret key called PEMK, respectively. As the EPs are on different devices they end up calculating different PEMKs, i.e., PEMK1 between the PaC and the Gateway, PEMK2 between the Relay and the actual Querier. One way of calculating this key [18] is:

```
PEMK = prf+(MSK, "IETF PEMK"|
            SID|KID|EPID)
```

Here, prf+ is a pseudo-random function defined in [16]. "IETF PEMK" is the ASCII code representation, SID is a four-octet Session Identifier, KID is associated with the MSK and EPID is the identifier of the EP. This key is specific to the multicast group that the EU has joined at the application level, and will be used for authorization at the network layer.

With those PEMKs, they establish a two different IPsec GSAs between them for cryptographic protection of IGMP messages. Each IPsec GSA contains one GSA_r and one GSA_q (for details see Section IV-B2). This phase is also used to test liveness of the PANA session.

4) Re-authentication and Termination phases are similar to that described in [14], except the fact that these PANA messages are exchanged through the AMT Tunnel.

2) Access Control at Network Level: In SIGMP [8], some messages are protected by IPsec GSAs. In this protocol all the operations for OGQ (Open Group Query) and OGR (Open Group Report) are retained from IGMPv3. However, for the access control of secure groups, a few operations are added in it. The material below describes how SIGMP is fitted into AMT.

- EU Operations: Once the Authentication is done at the application level, the EU will make his/her request for the network-level join and will send an SIGMP report message, believing it is being sent to the real Q. (In fact, it will be received by the SIGMP proxy in the Gateway). If this is the first time, when the report is sent to the IPsec (GSA) module, GSAM will be invoked to negotiate the cryptographic parameters (keys and SPIs) (see bullet 3, below). The IPsec module will then be able to send the report protected by those secure parameters to the Gateway where the SIGMP proxy (q) is implemented. The q in the Gateway will forward the message to the Relay through a secured tunnel (assumed) and finally the Relay will forward it to the actual Q that accepts the request.
- Q Operations: On receiving a secured report, Q will invoke IPsec module to decrypt it.
- GSAM: Group Security Association Management Protocol (GSAM) manages IPsec GSAs in two phases. In phase1, mutual authentication of EU and Q is done to achieve the registration of an EU. In phase 2, Q creates and distributes a GSA pair (GSA_q, GSA_r), named GSAM-TEK-SA to protect SIGMP messages (for details see [8]). Usually, in an IP multicast environment, GSAM negotiations are done between the EU and the real Q, but in AMT we must not let the EU communicate directly with the actual Q. As explained earlier, we implement an SIGMP proxy, which acts as querier functionality (q) in the Gateway, so that EU starts mutual authentication with the Gateway (q) using the derived PANA secret key, i.e., PEMK1. After authentication is done the Gateway (q) creates and distributes GSAM-TEK-SA (SA pair) to EU. On the other side of AMT tunnel SIGMP proxy acting as EU in the Relay performs mutual authentication with the actual Querier (Q) using PEMK2 and receives a GSA pair from Q.

V. Alternate Solutions

To our knowledge, the only other solution to providing access control for multicast services is based on having access control lists, either in the Set Top Box (STB) adjacent to the customer equipment, or in the access router. These solutions assume that the ISP has strong control over the STB or the access router. Our solution makes it possible for control to be exercised within the software of the Gateway. The existence of the "ticket", and the keys derived from the information in the ticket, ensure that the GO retains control of the session, in spite of the fact that the Gateway software would be freely available for downloading by the subscribers.

 ISBN: 978-1-61208-398-8

VI. AVISPA

The Automated Validation of Internet Security Protocols and Applications (AVISPA) project [5] has built a suite of tools that provides a modular and expressive formal language (High Level Protocol Specification Language, HLPSL) for specifying protocols and their security properties, and integrates different back-ends that implement a variety of automatic protocol analysis techniques. Experimental results, carried out on a large library of Internet security protocols, indicate that the AVISPA Tool is a state-of-the-art tool for Internet security protocol analysis as, to our knowledge, no other tool exhibits the same level of scope and robustness while enjoying the same performance and scalability [5]. In this section, we describe how we have transformed our model into HLPSL code, and how we have formulated the security goals to achieve the desired validation of the protocol.

- Our model in HLPSL code has four basic roles. They are client, server, gateway, relay. (Roles in AVISPA begin with a lower-case letter.) The roles client and server serve as PaC and PAA, respectively. As per our model, gateway and relay are acting as a friendly Man-in-the-Middle; they form SAs with client and server, respectively, and forward the EAP/PANA messages accordingly. The roles of the gateway and the relay are important because attacks are possible on both the gateway and the relay. So, we consider all four roles as main actors in HLPSL.
- In the real world, there is a large number of clients asking for a specific multicast application and they may request different multicast data streams as well. So, there is a need to distinguish all these clients and their requests. For that reason, we have added constants such as request-id and response-id, which assign a random unique number for each request made by clients. We transferred these constants along with nonces of client and server in initial request messages.
- After a few initial messages, PANA starts carrying EAP method (EAP-FAST) for authentication. EAP-FAST is already validated between two nodes in [13]. Now we implement it among four nodes in our HLPSL code. As phase 1 in EAP-FAST results in a shared key (SA) between two nodes, to make it simpler we introduced a shared key K1 between client, gateway and K2 between relay and server. Client and gateway protect further data with K1 and relay and server with K2.
- After authentication all the four roles are able to calculate a secret key (MSK). Using MSK and PANA nonces, they calculate MAC (Message Authentication Code) value as well. Our goal is to maintain the secrecy of secret keys MSK, K1, K2. Derivation of secret key (MSK) and MAC is shown in Figure 5 below.
- After calculation of above mentioned keys, the results are passed to client from server.
- The session role defines executing of several basic roles in parallel. In our HLPSL code, the session role is composed of client, gateway, relay and server roles. Every role has two channels, send and receive, on which they send and receive messages. We should run these four roles in parallel for messages to pass through the AMT tunnel (see Figure 6).

```
% Calculation of Master Session Key.
Msk' := H(Nec'.Nes.Psk')

% Calculation of Message Authentication Code
Mac' := INT(PRF(H(Nec'.Nes.Psk').Nps.Npc.
Kid).Pmsg)
```

Fig. 5. Secret Key and Message Authentication Code.

```
role session(
C,G,R,S :agent
K1,K2 :symmetric
H,PRF,INT :hash_func)
def=
local SC,RC,SG,RG,SR,RR,SS,RS :channel (dy)
composition
client (C,G,R,S,K1,H,PRF,INT,SC,RC)
/\ gateway(C,G,R,S,K1,H,PRF,INT,SG,RG)
/\ relay (C,G,R,S,K2,H,PRF,INT,SR,RR)
/\ server (C,G,R,S,K2,H,PRF,SS,RS)
end role
```

Fig. 6. Session Role.

- In the environment role, we can modify the number of parallel sessions and the knowledge of intruder. In our code, the intruder has been given the knowledge of all the hash functions, agents and his own private key. First, we executed a session without any intruder. In the next step, we executed session with client as intruder and then gateway, relay, server as intruders (see Figure 7).

```
role environment()
def=
const C,G,R,S :agent,
KK1,KK2 :protocol_id,
K1,K2,Ki :symmetric_key, H,PRF,INT
:hash_func
intruder_knowledge = {c,g,r,s,h,prf,int,ki}

composition
session(c,g,r,s,k1,k2,h,prf,int)
/\ session(i,g,r,s,ki,k2,h,prf,int)
/\ session(c,i,r,s,ki,k2,h,prf,int)
/\ session(c,g,i,s,k1,ki,h,prf,int)
/\ session(c,g,r,i,k1,ki,h,prf,int)
end role
```

Fig. 7. Environment Role.

- In the goal section of our HLPSL code, we explicitly ask the AVISPA model checker to validate the secrecy of both the shared secret keys (K1, K2) and MSK, which ensures the intended security of further communications. Security goals are shown in Figure 8.
- Considering the security goals mentioned in the goal section of our HLPSL code, no attack has been found.

 ISBN: 978-1-61208-398-8

```
goal
%Secrecy of Shared Key between Client
% and Gateway
secrecy_of kk1

%Secrecy of Shared Key between Relay
% and Server
secrecy_of kk2

%Secrecy of Master Session Key)
secrecy_of s_msk
end goal
```

Fig. 8. Goals.

Summary results of three AVISPA back-ends OFMC, CL-AtSe and SATMC appeared to be safe. This shows our model (Receiver Access Control in AMT) in reality is immune to all those potential attacks and threats.

VII. Conclusion

In this paper, we have proposed a solution that provides receiver access control for multicast groups in the AMT environment. This solution allows only legitimate End Users in a unicast-only-network to access networks and receive multicast data from multicast enabled sources. We have used AVISPA to formally demonstrate the security of these extensions to AMT.

Acknowledgment

J. W. Atwood acknowledges the support of the Natural Sciences and Engineering Research Council of Canada, through its Discovery Grants program.

References

[1] J. W. Atwood, "An architecture for secure and accountable multicasting," in *32nd IEEE Conference on Local Computer Networks (LCN 2007)*, Oct. 2007, pp. 73–78.

[2] S. Islam and J. W. Atwood, "Sender access and data distribution control for inter-domain multicast groups," *Computer Networks*, vol. 54, no. 10, pp. 1646–1671, 2010.

[3] ——, "Multicast receiver access control by igmp-ac," *Computer Networks*, vol. 53, no. 7, pp. 989–1013, 2009.

[4] ——, "Multicast receiver access control using pana," in *Proceedings of the 1st Taibah University International Conference on Computing and Information Technology (ICCTT 2012)*, Mar. 2012.

[5] L. Viganò, "Automated security protocol analysis with the AVISPA tool," *Electronic Notes in Theoretical Computer Science*, vol. 155, no. 10, pp. 61–86, 2006.

[6] S. Islam, "Participant access control in ip multicasting," Ph.D. dissertation, Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada, Mar. 2012.

[7] S. Kent and K. Seo, "Security architecture for the internet protocol," Internet Engineering Task Force, Request for Comments 4301, Dec. 2005, URL: http://www.rfc-editor.org/rfc/rfc4301.txt [accessed: 2015-02-15].

[8] B. Li and J. W. Atwood, "Receiver access control for IP multicast at the network level," *Submitted to Computer Networks*, Sep. 2014.

[9] B. Cain, S. Deering, I. Kouvelas, B. Fenner, and A. Thyagarajan, "Internet group management protocol, version 3," Internet Engineering Task Force (IETF), Request for Comments 3376, Oct. 2002, URL: http://www.rfc-editor.org/rfc/rfc3376.txt [accessed: 2015-02-15].

[10] B. Fenner, M. Handley, H. Holbrook, and I. Kouvelas, "Protocol independent multicast - sparse mode (pim-sm): Protocol specification (revised)," Internet Engineering Task Force (IETF), Request for Comments 4601, Aug. 2006, URL: http://www.rfc-editor.org/rfc/rfc4601.txt [accessed: 2015-02-15].

[11] V. Fajardo, J. Arkko, J. Loughney, and G. Zorn, "Diameter base protocol," Internet Engineering Task Force, Request for Comments 6733, Oct. 2012, URL: http://www.rfc-editor.org/rfc/rfc6733.txt [accessed: 2015-02-15].

[12] B. Aboba, L. Blunk, J. Vollbrecht, J. Carlson, and H. Levkowetz, "Extensible authentication protocol (eap)," Internet Engineering Task Force, Request for Comments 3748, Jun. 2004, URL: http://www.rfc-editor.org/rfc/rfc3748.txt [accessed: 2015-02-15].

[13] M. Parham, "Validation of the security of participant control exchanges in secure multicast content delivery," Master's thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada, Sep. 2011.

[14] D. Forsberg, Y. Ohba, B. Patil, H. Tschofenig, and A. Yegin, "Protocol for carrying authentication for network access (pana)," Internet Engineering Task Force, Request for Comments 5191, May 2008, URL: http://www.rfc-editor.org/rfc/rfc5191.txt [accessed: 2015-02-15].

[15] P. Jayaraman, R. Lopez, Y. Ohba, M. Parthasarathy, and A. Yegin, "Protocol for carrying authentication for network access (pana) framework," Internet Engineering Task Force, Request for Comments 5193, May 2008, URL: http://www.rfc-editor.org/rfc/rfc5193.txt [accessed: 2015-02-15].

[16] Y. Nir, C. Kaufman, P. Hoffman, and P. Eronen, "Internet key exchange (ikev2) protocol," Internet Engineering Task Force, Request for Comments 5996, Sep. 2010, URL: http://www.rfc-editor.org/rfc/rfc5996.txt [accessed: 2015-02-15].

[17] M. Parthasarathy, "Pana enabling ipsec based access control," Internet Engineering Task Force, Internet Draft, Work in progress, Dec. 2005, URL: http://tools.ietf.org/id/draft-ietf-pana-ipsec-07.txt [accessed: 2015-02-15].

[18] Y. Ohba and A. Yegin, "Definition of master key between pana client and enforcement point," Internet Engineering Task Force, Request for Comments 5807, Mar. 2010, URL: http://www.rfc-editor.org/rfc/rfc5807.txt [accessed: 2015-02-15].

[19] B. Weis, G. Gross, and D. Ignjatic, "Multicast extensions to the security architecture for the internet protocol," Internet Engineering Task Force, Request for Comments 5374, Nov. 2008, URL: http://www.rfc-editor.org/rfc/rfc5374.txt [accessed: 2015-02-15].

[20] G. Bumgardner, "Automatic multicast tunneling," Internet Engineering Task Force, Request for Comments 7450, Feb. 2015, URL: http://www.rfc-editor.org/rfc/rfc7450.txt [accessed: 2015-03-02].

[21] L. Morand, A. Yegin, S. Kumar, and S. Madanapalli, "Dhcp options for protocol for carrying authentication for network access (pana) authentication agents," Internet Engineering Task Force (IETF), Request for Comments 5192, May 2008, URL: http://www.rfc-editor.org/rfc/rfc5192.txt [accessed: 2015-02-15].

[22] V. N. T. Malla, "Design and validation of receiver access control in the automatic multicast tunneling environment," Master's thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada, Aug. 2014.

[23] N. Cam-Winget, D. McGrew, J. Salowey, and H. Zhou, "The flexible authentication via secure tunneling extensible authentication protocol method (eap-fast)," Internet Engineering Task Force (IETF), Request for Comments 4851, May 2007, URL: http://www.rfc-editor.org/rfc/rfc4851.txt [accessed: 2015-02-15].

[24] A. Yegin, Y. Ohba, R. Penno, and C. Wang, "Protocol for carrying authentication for network access (pana) requirements," Internet Engineering Task Force, Request for Comments 4058, May 2005, URL: http://www.rfc-editor.org/rfc/rfc4058.txt [accessed: 2015-02-15].

 ISBN: 978-1-61208-398-8

Benchmarking the Performance of XenDesktop Virtual DeskTop Infrastructure (VDI) Platform

Shie-Yuan Wang

Department of Computer Science
National Chiao Tung University, Taiwan
Email: shieyuan@cs.nctu.edu.tw

Wen-Jhe Chang

Department of Computer Science
National Chiao Tung University, Taiwan
Email: ethan-shy@hotmail.com

***Abstract*—The recent advances in portable devices and the trends to move a user's desktop to cloud environments have changed how people use traditional computers today. Several companies have developed the "Virtual Desktop Infrastructure" (VDI) technology for this trend. By this technology, people need not use a traditional PC with a high clock-rate CPU and a large storage device to run heavy tasks. Instead, they can use a "thin-client" and a network connection to run these heavy tasks on a remote VDI server. A VDI user can perform these tasks in a virtual desktop run by a virtual machine (VM) on the VDI server. During operations, the screen image of the virtual desktop will be delivered to the screen of the thin-client. The VDI technology offers many advantages. However, it may increase the perceived delays when a VDI user operates a virtual desktop. These delays may be caused by bad network conditions or by overloading conditions on the VDI server. In this paper, we developed a VDI performance benchmarking tool and used it to measure the perceived delays when the XenDesktop VDI platform is used under various network conditions and overloading conditions on the VDI server. The EstiNet network simulator and emulator was used to create various network conditions for benchmarking measurements.**

Keywords–EstiNet; VDI; VM.

I. INTRODUCTION

Recently, the VDI technology has become more and more popular due to the availability of high-speed network accesses and advances in portable devices. In this architecture, users operate their virtualized desktops on a remote VDI server rather than operate a real desktop on their local machines. When they operate from their local machines, all of the programs, applications, processes, and data are run and kept on the VDI server. In this way, a user can run the same operating system and execute the same applications and access the same data from any machine via a network connection. Because this computing model has great potential to save cost and increase data security, many companies such as Citrix [1], Microsoft [2], Oracle [3], and VMware [4] have developed their own VDI technologies.

VDI offers many advantages but also comes with several challenges. Nowadays, many users still hesitate to adopt the VDI technology. One major concern is that using VDI may suffer a much higher delay than using a desktop computer and it is difficult for the VDI user to find out the causes. VDI is a client-server architecture. When VDI users operate their virtual desktops through a network, they must compete with other VDI users for the network bandwidth and the various resources on the VDI server. As a result, many factors can cause the VDI users to more easily experience long delays when operating their remote virtual desktops. Besides, because a virtual desktop is run by a VM on a remote VDI server, it is more difficult for the VDI users to find out what factor is causing the long operation delays. For example, either a high CPU usage on the VDI server or a long round-trip network delay between the VDI user and the VDI server can cause the user to experience large delays. However, the VDI user does not know which factor is causing this delay and thus does not know whether he should contact the VDI cloud service provider or the Internet service provider to report and complain the bad performance problems.

In this paper, we develop a performance benchmarking tool to measure the delay (i.e., the screen response time) of Citrix's XenDesktop VDI platform under five important conditions. The first two conditions: (1) large link delay and (2) high packet loss rate, are about the quality of the network. The other three conditions: (3) high CPU usage, (4) insufficient memory allocation per VM, and (5) high disk usage, are about the VDI server resource usage conditions. According to our measured results, each of these five factors can cause a large delay when a VDI user performs tasks on his virtual desktop. Our results reveal that these factors affect the delay differently. Due to the paper length limitation, in this paper we can only present the performance benchmarking results. In our future paper, we will present how we use the delay features of these factors to develop a VDI performance diagnostic tool that can accurately tell a VDI user which factor(s) is (are) causing the long perceived delay.

The rest of the paper is organized as follows. In Section II, we present related work on virtual desktop infrastructure. In Section III, we present the implementation of our performance benchmarking tool. Experimental setups are presented in Section IV and various experimental results are presented in Section V. Finally, we conclude the paper in Section VI.

II. RELATED WORK

With the advances of desktop virtualization and thin-client devices, there have been some work on performance evaluation of VDI. The closest work to ours is DeskBench [5], which captures the screen and records and playbacks keyboard and mouse events on the client side. The other close work is VNCPlay [6], which is also based on matching screen and recording and playback of keyboard and mouse events. Another similar work is Slow Motion Benchmarking [7]. It captures the network traffic exchanged between the client and server and replays the network traffic later in slow motion. The

 ISBN: 978-1-61208-398-8

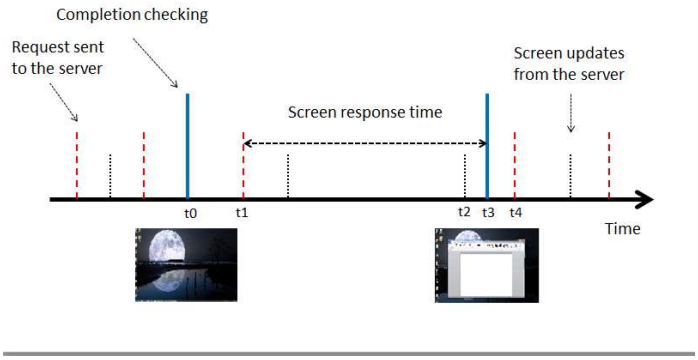


Figure 1. Interaction between the VDI client and server

authors in [8] proposed the VDBench toolkit to measure thin-client performances under server load and network conditions. In [9], the authors focused on benchmarking the audio transmission performance when virtual desktop platform is used, as this is the topic that a Telcom operator would concern.

In our paper, we use a similar approach to these work to measure the screen response time. However, we use the Citrix's XenDesktop VDI platform as the performance study target while most of these previous work studied the open-source VNC protocol and Microsoft's Remote Desktop Protocol (RDP). Compared with the work done in [8], the authors in [8] used VMware ESXi 4.0 as the VDI server while we used Citrix's XenDesktop VDI platform as the VDI server. XenDesktop uses a patented highly-efficient HDX technology as the VDI protocol between its VDI client and server. According to our measurements and comparisons, we found that HDX performs much better than VNC and RDP because the perceived VDI delay when HDX is used is much less than those when VNC and RDP are used. Beside the differences in the tested VDI technologies, in this paper we focus on the XenDesktop delay performances under various network conditions and overloading conditions on the VDI server. Most of these conditions are not studied in these previous work.

III. Implementation of the Performance Benchmarking Tool

The screen response time of a VDI user's action is the time between when the user clicks the mouse or does the keyboard input and the time when the corresponding screen shows up completely on the VDI user's device. The response time measuring process is depicted in Figure 1. The horizontal line represents the time axis. Vertical dashed lines represent the requests sent from the client to the server. Short vertical dotted lines represent screen updates arriving from the sever. High vertical solid lines represent when our tool compares the current screen image with the expected one. Assume that the user clicks the mouse left button to execute Microsoft Word application at time t1. Further assume that the server sends the new screen of executing Microsoft Word to the client at t2 and our tool detects that a match occurs at t3. For this example case, the time between when the user sends a request to the server and the time when the corresponding screen shows up is t3 - t1. This is the VDI delay performance measured and reported in this paper.

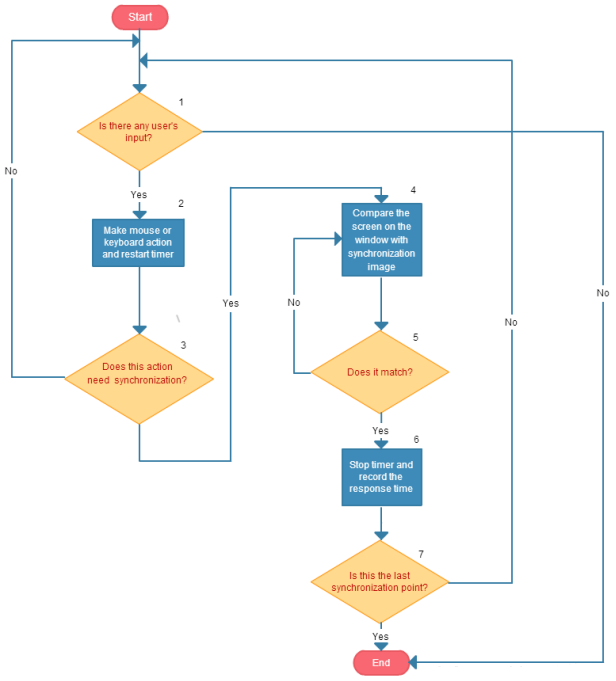


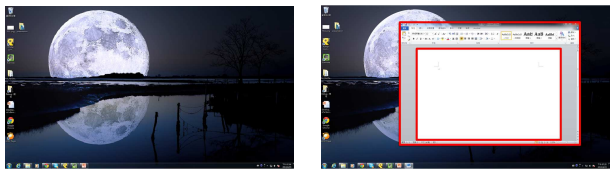
Figure 2. Flow chart of measuring screen response time

The flow chart illustrating how we automatically generate a user's action and measure its screen response time is shown in Figure 2.

1 First, check if there is any user's input occurring.
2 If input occurring, our tool uses the WIN32 API to capture and replay the user's actions of mouse and keyboard.
3 If an action needs to synchronize with the screen (that is, after executing the action, a new screen must show up before the next action can be executed), our tool will start a timer to record the response time. (Note that some actions need not synchronize with the screen. In such a case, the next action in the replay list can be executed immediately without waiting for the new screen to show up.)
4 For an action that needs screen synchronization, after the timer is started, our tool will periodically check whether the expected screen image has arrived from the server by comparing the current screen image of an area with the new screen image of the same area.
5 Detect if there is any screen match occurring.
6 If a screen match is detected, our tool will stop the timer and record the response time of this action. Then, the tool will execute the next action.
7 If there are no more actions to send to the server, the tool will stop and show the response time of all executed actions.

To make the measured response time accurate, when the timer is turned on, the interval between two successive screen image comparisons must be small enough. This means that each image comparison must be finished as soon as possible. To do so, our tool intelligently compares only the new and old images of the area where its screen image is expected to change without comparing the new and old screen images of the whole screen. As an example, Figure 3(a) shows the whole

 ISBN: 978-1-61208-398-8



(a) Screen before an action is sent (b) Screen after an action is completed

Figure 3. Screen before and after a user sends a request

screen before an action is sent to the server while Figure 3(b) shows that a window is popped up after the action is sent to the server and completed. We see that this action only changes an area of the screen. As a result, our tool only needs to compare an area of the whole screen to speed up the image comparison operation. With this improvement, our tool can finish the image comparison in 100 microseconds. In our implementation, when the measuring timer is turned on, our tool will perform the image comparison every 100 microseconds.

The three types of actions that most users will issue when performing tasks on a desktop are listed in Table I. Because these types of actions will be executed very frequently, the VDI delays of these actions are important performance metrics that can be used to judge the delay quality experienced by a VDI user.

TABLE I. TYPES OF ACTIONS ISSUED ON THE CLIENT TO THE VDI SERVER

Action Types	Explanation
Opening and closing Microsoft Word	The client sends the mouse click action to the server to open/close the Microsoft Word application. It then measures the screen response time for the application to completely open up/close down its window.
Keyboard input	The client sends the keyboard input action to type some few words in Microsoft Word. It then measures the screen response time of the words showing up on the screen.
Compressing files	The client sends the mouse click action to perform a compress files operation. It then measures the screen response time of the compressing application finishing its jobs completely.

IV. EXPERIMENTAL SETUP

In this paper, we use Citrix's XenDesktop as our VDI platform and use the EstiNet network simulator and emulator [10]–[12] to create various network conditions between the VDI client and server. The setup of our experiments is presented in Figure 4. The lower part of Figure 4 shows that in the real world we use Asus RS500A-E6 for our VDI server, which is equipped with two AMD CPUs (each with 12 cores operating at 1.9 GHZ) and 48 Gigabyte memory. The VDI server runs XenServer (version 5.6) to host up to 32 VMs each running Windows 7 operating system. It also runs the XenDesktop controller to manage these Windows 7 virtual desktops. The VDI client runs XenDesktop receiver and our tool on a PC that is equipped with an Intel CPU (3.4 GHZ dual core). We run EstiNet on another computer as a network emulator between the VDI client and server. It connects to both the VDI client and server to intercept their exchanged packets to vary the delay and packet loss rates experienced by these packets.

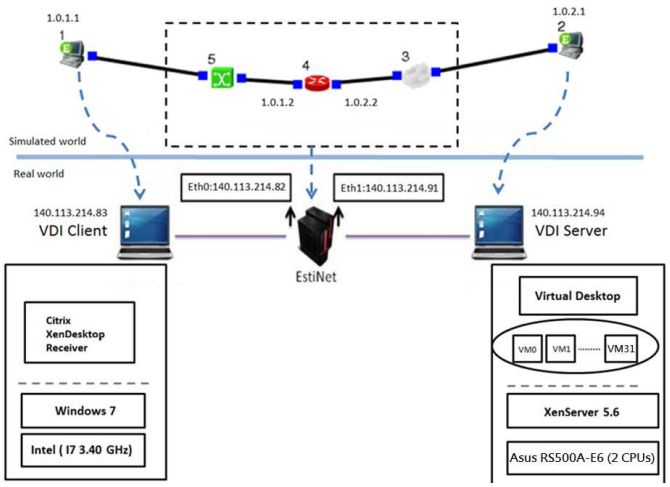


Figure 4. Experimental Environment

The upper part of Figure 4 shows how the tested network in the real world is represented and simulated in the EstiNet network emulator. In the simulated network, node 1 with the IP address 1.0.1.1 represents the VDI client while node 2 with the IP address IP 1.0.2.1 represents the VDI server. Because the packets exchanged between node 1 and node 2 will go through node 3, which is configured to add certain delays to these packets or drop these packets at a certain rate, EstiNet can easily vary the network conditions between the VDI client and server in the real world.

V. EXPERIMENTAL RESULTS

We measure the response time of opening and closing Microsoft Word, keyboard input, and compressing files under different server loadings, link delays, and packet loss rates. We vary the server's total CPU usage by running 0-32 VMs on the server and let each VM run a CPU-bound job, which consumes about one CPU core. We vary the size of the memory allocated to each VM from 4 GB down to 1 GB and execute a program on each VM to purposely consume about 1 GB memory. Doing so is to test how important the usable memory space is to the VDI delay of an action. We also vary the disk usage on the server by running 0-10 VMs on the server and let each VM compress files to generate about 6 MB/s disk read/write load per VM. We found that 10 VMs are enough to generate heavy loads on the disk. In the following figures, each data point is the average of 100 runs of the measurements of the same type of action performed under the same settings.

Figures 5 - 7 show the response time of opening and closing Microsoft Word under different server loading conditions. A first finding is that the delay of the "Word Close" action is much less than the delay of the "Word Open" action, and its delay remains low and stable under high server loading conditions. These results suggest that the "Word Close" action is a light-weight operation. In contrast, from Figure 5 we see that when more and more CPU-bound VMs are competing for the shared CPU resource, which results in insufficient CPU resource allocation for the VM executing the "Word Open" action, the delay of this action goes up quickly. From Figure 6, we see that when the allocated memory to a VM is less than 1.5 GB due to the competition among more and more VMs, the

 ISBN: 978-1-61208-398-8

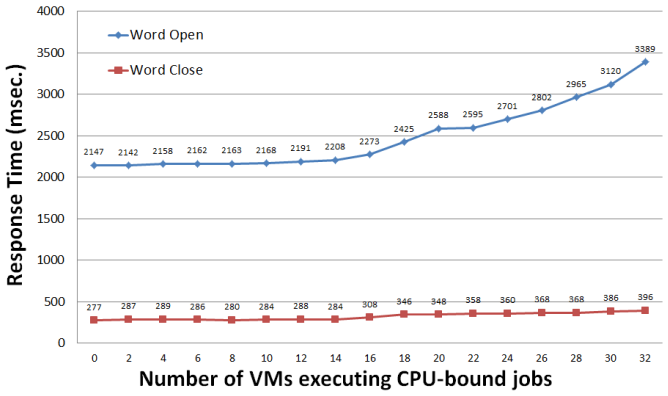


Figure 5. Screen response time of opening and closing Microsoft Word under different numbers of VMs each executing CPU-bound jobs

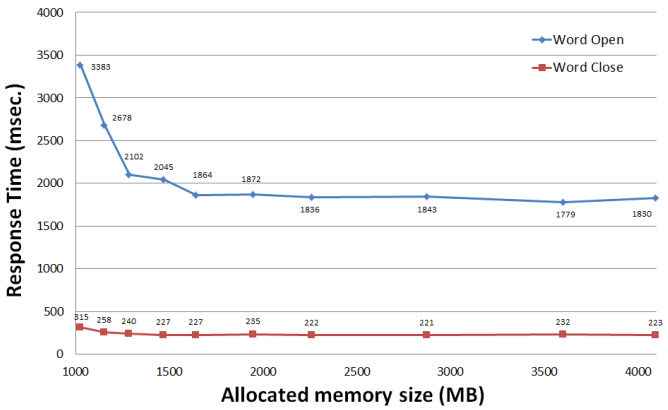


Figure 6. Screen response time of opening and closing Microsoft Word under different allocated memory sizes per VM

delay of the "Word Open" action goes up very quickly. This phenomenon is due to the "thrashing" effect of the operating system, which refers to the situation when the hard disk space is used as memory space due to insufficient memory space allocated for executing an application.) From this phenomenon, we see that the "Word Open" action not only requires much CPU resource, it also requires much memory resource for quick response. Figure 7 shows that even with high disk usages (which is caused by compressing applications executing many disk I/O operations), the delay of "Word Open" action does not increase much. This suggests that the "Word Open" action does not require much disk throughput resource. In summary, our results show that the VDI delay of "Word Open" action increases under high CPU usage or high memory usage but remains about the same under high disk usage.

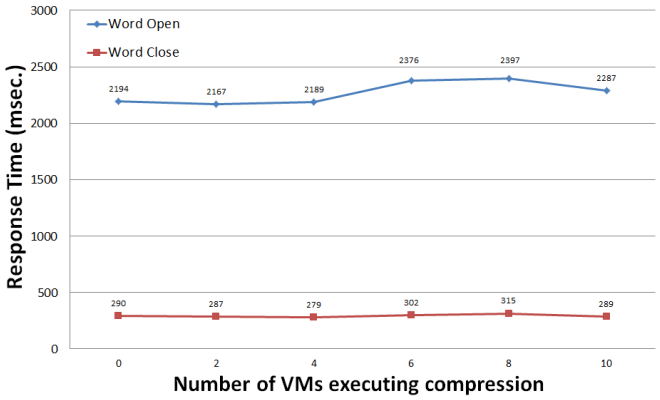


Figure 7. Screen response time of opening and closing Microsoft Word under different numbers of VMs each executing compression

The delays of the "Word Open" action under different link delays are shown in Figure 8. As expected, the link delay linearly affects the delay of the action because the VDI client must wait for the link delay to elapse before the screen update can arrive. Figure 9 shows the delay of the "Word Open" action under different packet loss rates. We see that the packet loss rate increases the delay non-linearly and when the packet loss rate exceeds 12%, the delay starts to increase dramatically. This phenomenon can be explained by the fact that the transport protocol used by XenDesktop's VDI technology is TCP and TCP throughput is very sensitive to the packet loss rate due to its conservative congestion control algorithm.

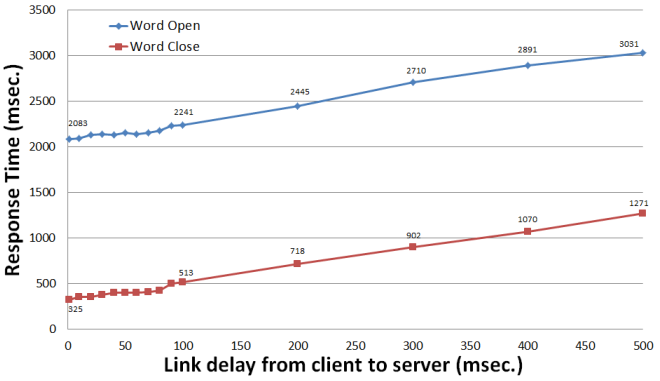


Figure 8. Screen response time of opening and closing Microsoft Word under different link delays

Figures 10 - 14 show the response time of keyboard input under different server loading and network conditions. Figure 10 shows that the "Keyboard input" action requires some CPU resource and its delay increases by a small amount of 30 milliseconds when more and more VMs are competing for the shared CPU resource. Figure 11 shows that the delay of the "Keyboard input" action remains about the same under different allocated memory sizes unless the size of the allocation drops to 1 GB, where the thrashing effect starts to begin. Figure 12 shows that the delay of the "Keyboard input" action does not increase as the disk usage increases. This phenomenon is expected as the processing of a keyboard input on the VDI server does not need to use any disk I/O operation. As a result, the delay of the "Keyboard input" action has no relationship with the current disk usage on the VDI server. From these three figures, we see that the maximum and minimum delays measured for the "Keyboard input" action under different server loading conditions only differ by about 30 milliseconds. This difference is quite small and the VDI user will not notice such a difference.

However, Figure 13 and Figure 14 show that the network conditions can dramatically increase the delay of keyboard input. As expected, the link delay between the VDI client

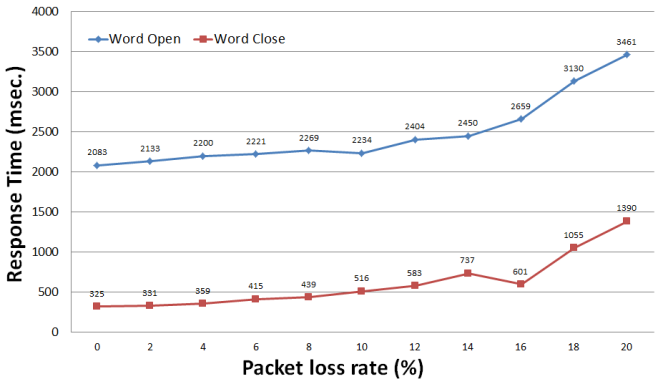


Figure 9. Screen response time of opening and closing Microsoft Word under different packet loss rates

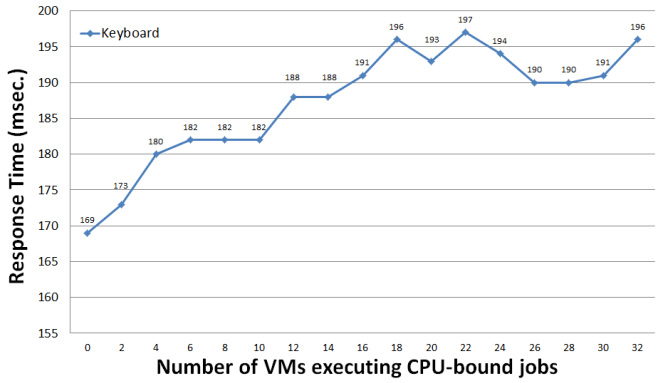


Figure 10. Screen response time of keyboard input under different numbers of VMs each executing CPU-bound jobs

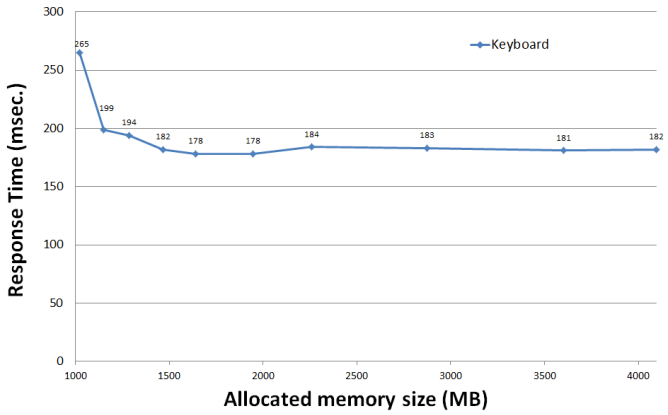


Figure 11. Screen response time of keyboard input under different allocated memory sizes per VM

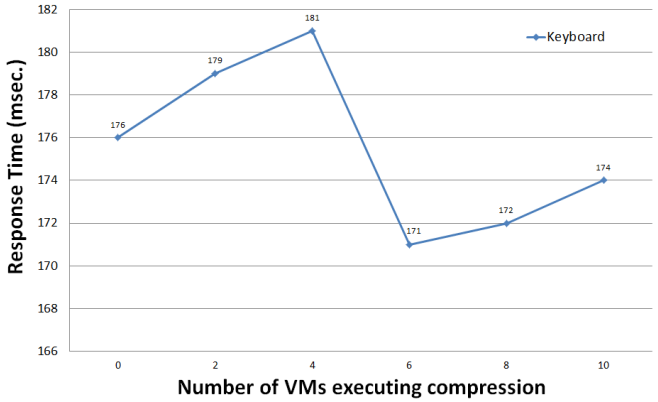


Figure 12. Screen response time of keyboard input under different numbers of VMs each executing compression

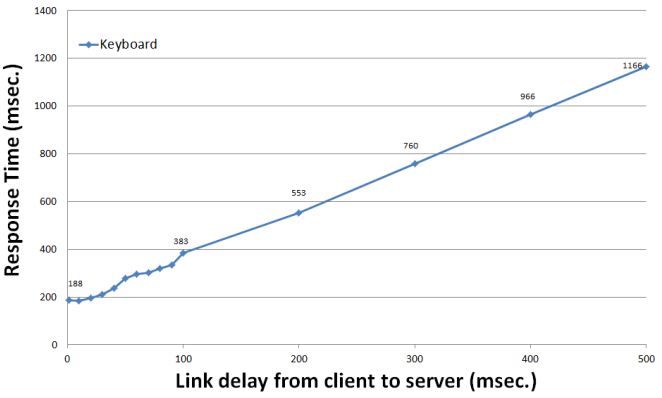


Figure 13. Screen response time of keyboard input under different link delays

and server linearly increases the delay of keyboard input. This is reasonable as the data amount that needs to be exchanged between the VDI client and server for a keyboard input is only a few bytes. Thus, the processing time required for the exchanged data on the VDI server is little and constitutes only a very tiny portion of the delay. That is, most of the delay comes from the link delay between the VDI client and server. Regarding the packet loss rate, we see that it also affects the delay of keyboard input significantly. We see that when the packet loss rate exceeds 18%, the delay abruptly jumps to 1.078 seconds, which is very noticable and annoying for the VDI user.

In summary, we found that the delay of the "Keyboard input" action generally is not affected by the server loading conditions but will be affected by large link delays and high packet loss rates in the network.

Figure 15 - 17 show the response time of compressing files under different server loading conditions. We do not measure the response time of compressing files under network conditions. This is because the time required to finish compressing files is too large (e.g., above 100,000 ms) compared to the tested link delays, whose maximum value is 500 ms. For such a situation, the link delay affects the delay of compressing files very minimally. In addition, because the "compressing files" action compresses the files on the VDI server without the need to transfer any file from the VDI client to the server, the packet loss rate does not affect the delay of the "compressing files" action at all. As a result, we do not study its effects on the delay of the "compressing files" action.

As shown in Figure 15, the "compressing files" action requires much CPU resource. This is evident because the figure shows that when more and more CPU-bound VMs are competing for the CPU resource, the delay of the "compressing files" action increases rapidly. Figure 16 shows that the "compressing

 ISBN: 978-1-61208-398-8

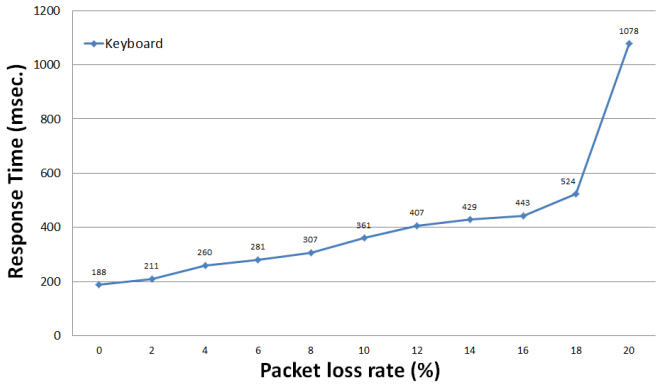


Figure 14. Screen response time of keyboard input under different packet loss rates

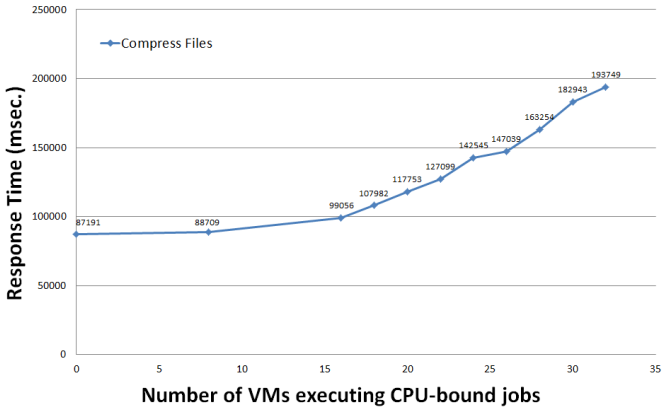


Figure 15. Screen response time of compressing files under different numbers of VMs each executing CPU-bound jobs

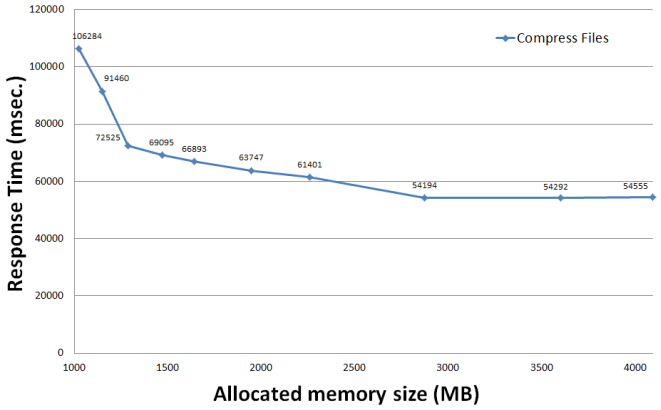


Figure 16. Screen response time of compressing files under different allocated memory sizes per VM

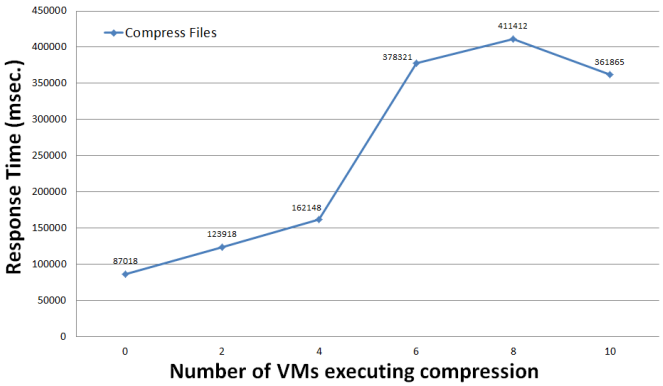


Figure 17. Screen response time of compressing files under different numbers of VMs each executing compression

files" action also requires much memory resource to finish the job quickly. Comparing Figure 16 with Figure 6 and Figure 11, we see that the "compressing files" action requires much more memory than the "Word Open" action or the "Keyboard input" action because its delay starts to go up when the allocated memory size is less than 3 GB while the delays of the other two actions go up only when the allocated memory sizes are less than 1.7 GB and 1.3 GB, respectively. Figure 17 shows that the "compressing files" action also requires much disk throughput resource because when more and more VMs are compressing files to compete for the disk throughput resource, the delay of the "compressing files" action goes up rapidly. From these figures, we see that the delay of the "compressing files" action is affected by high CPU usage, insufficient memory allocation, and high disk usage.

VI. Conclusion

In this paper, we developed a VDI performance benchmarking tool and used it to study the delay performance of the XenDesktop VDI platform under various network conditions and server loading conditions. Using the EstiNet network emulator to create various network conditions, our study shows that the following factors can increase the experienced delay of XenDesktop VDI significantly: the link delay and the network packet loss rate between the VDI client and server, the CPU utilization of the VDI server, the disk read/write load of the VDI server, and the size of the memory allocated to a VM running the virtual desktop. Our measurement results reveal that these factors affect the experienced delay of XenDesktop VDI differently. The delay features of these factors can be used to judge what factor(s) is (are) causing the delay when a user operates a virtual desktop. Based on these unique delay features, a diagnostic tool can be developed to help network service providers and cloud service providers to jointly identify the real causes for large experienced VDI delays. In the future, we will extend our work to study the perceived VDI delays when the VM that runs a virtual desktop migrates from one physical server to another. This topic is important as a VM may frequently migrate in a cloud for load balancing purposes.

References

[1] "XenDesktop Product Information," URL: http://www.citrix.com [accessed: 2015-03-04].

[2] "Virtual Desktop Infrastructure in Windows Server 2008," URL: http://www.microsoft.com [accessed: 2015-03-04].

[3] "Oracle Virtual Desktop Infrastructure Product Information," URL: http://www.oracle.com [accessed: 2015-03-04].

[4] "VMWare EMC." URL: http://www.vmware.com [accessed: 2015-03-04].

[5] J. Rhee, A. Kochut, and K. Beaty, "Deskbench: flexible virtual desktop benchmarking toolkit," in Integrated Network Management, 2009. IM'09. IFIP/IEEE International Symposium on. IEEE, 2009, pp. 622–629.

 ISBN: 978-1-61208-398-8

[6] N. Zeldovich and R. Chandra, “Interactive performance measurement with vncplay.” in USENIX Annual Technical Conference, FREENIX Track, 2005, pp. 189–198.

[7] J. Nieh, S. J. Yang, and N. Novik, “Measuring thin-client performance using slow-motion benchmarking,” ACM Transactions on Computer Systems (TOCS), vol. 21, no. 1, 2003, pp. 87–115.

[8] A. Berryman, P. Calyam, M. Honigford, and A. M. Lai, “Vdbench: A benchmarking toolkit for thin-client based virtual desktop environments,” in Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on. IEEE, 2010, pp. 480–487.

[9] F. Wang, Y. Liu, B. Lei, and J. Li, “Benchmark driven virtual desktop planning: A case study from telecom operator,” in Proceedings of the 2012 International Conference on Cloud and Service Computing. IEEE Computer Society, 2012, pp. 204–211.

[10] S.-Y. Wang, C.-L. Chou, and C.-M. Yang, “EstiNet OpenFlow network simulator and emulator,” Communications Magazine, IEEE, vol. 51, no. 9, 2013, pp. 110–117.

[11] S.-Y. Wang, “Comparison of SDN OpenFlow network simulator and emulators: EstiNet vs. Mininet,” in Proceedings of the 2014 IEEE International Symposium on Computers and Communication (ISCC). IEEE, 2014, pp. 1–6.

[12] “EstiNet Network Simulator and Emulator,,” URL: http://www.estinet.com [accessed: 2015-03-04].

 ISBN: 978-1-61208-398-8

Worst Case Modeling of Aggregate Scheduling by Network Calculus

Ulrich Klehmet Rüdiger Berndt
Computer Networks and Communication Systems
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Email: {ulrich.klehmet, ruediger.berndt}@fau.de

***Abstract*—*Network Calculus* (NC) is a powerful mathematical theory for the performance evaluation of communication systems, since it allows to obtain worst-case performance measures. In communications system modeling, the NC theory is often used to determine *Quality of Service* (QoS) guarantees, for example of packet-switched systems. In networking systems, the aggregation of data flows plays an important role while modeling the multiplexing scheme. When the multiplexing order is not *First in, First out* (FIFO), the strictness of the service curve plays an important role. This article deals with problems that arise from the strictness requirement considering aggregate scheduling. The literature reports that the strictness of an aggregated service curve is a fundamental precondition to obtain the individual service curve for a single left-over flow when a node processes multiple input flows in a non-FIFO manner. In many publications, this important strictness property is assumed to be a feature of the service curve only. We will show that, in general, this assumption is not true. In most cases, only the concrete input flow in connection with the service curve allows to decide whether the service curve is strict or non-strict. However, the abstraction from a concrete input flow with an arrival curve as upper bound is not enough to determine the service curve's strictness. Therefore, to bypass the strict-non-strict problems, we devise theorems to gain guaranteed performance values for a left-over flow.**

Keywords*– *Worst-case Communication System Modeling; Network Calculus; Aggregate Scheduling; Strict-non-strict Service Curves; Backlogged Period

I. Introduction

For systems with hard real-time requirements, *timeliness* plays an important role. This *Quality of Service* (QoS) requirement can be found in various kinds of systems, including cars, airplanes, industrial networks, or power plants [1]. Analytical performance evaluation of such systems cannot be based on stochastic modeling, like traditional queuing theory, only. Since worst-case performance parameters like maximum delay of service times [2] are required, the knowledge of mean values is not sufficient. In other words, one needs a mathematical tool to calculate performance figures—in terms of bounding values—which are valid in any case. Such a tool is *Network Calculus* (NC), as a novel system theory for deterministic queuing systems [3] [4].

This article is structured as follows: In Section II, we describe the basic elements of NC. Section III introduces aggregate scheduling and the use of NC w.r.t. the analysis of individual flows as part of an aggregation. Section IV shows possibilities to overcome the problems described in Section III. Finally, in Section V we draw a conclusion.

II. Basic modeling elements of Network Calculus

The most important modeling elements of NC are given by *arrival curves*, *service curves*, and the *min-plus convolution*. Arrival and service curves are the basis for the computation of maximum deterministic boundary values, like backlog bounds, or delay bounds [4].

Definition 1 (Arrival curve): Let $\alpha(t)$ be a non-negative, non-decreasing function. Flow F with input $x(t)$ at time t is constrained by the arrival curve $\alpha(t)$ iff:

$$\forall\ 0 \leq s \leq t : x(t) - x(s) \leq \alpha(t-s) \tag{1}$$

Flow F is also called α-smooth.

Example 1: A commonly used arrival curve is the token bucket constraint:

$$\alpha_{r,b}(t) = \begin{cases} r \cdot t + b & \text{for } t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Figure 1 shows a token bucket arrival curve (dashed blue line) forming an upper limit for a traffic flow $x(t)$ with an average rate r and an instantaneous burst b. This means for $\Delta t := t-s$ and $\Delta t \to 0$, $\lim_{t\to s}\{x(t)-x(s)\} \leq \lim_{\Delta t\to 0}\{r\cdot\Delta t+b\} = b$.

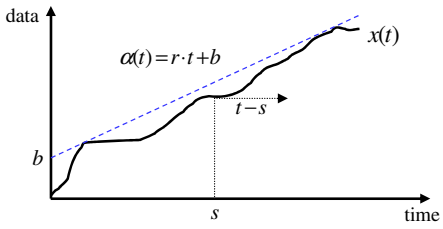


Figure 1. Token bucket arrival curve

Next to arrival curves, the convolution operation plays an important role in NC theory.

Definition 2 (Min-Plus Convolution): Let $f(t)$ and $g(t)$ be non-negative, non-decreasing functions that are 0 for $t \leq 0$. A third function, called *min-plus convolution* is defined as:

$$(f \otimes g)(t) = \inf_{0\leq s\leq t}\{f(s) + g(t-s)\} \tag{3}$$

By applying this operation, arrival curve $\alpha(t)$ can be characterized with respect to $x(t)$ as:

$$x(t) \leq (x \otimes \alpha)(t) \tag{4}$$

Service curves, in contrast to arrival curves, are used to model the output of a system—for example to determine whether there is a guaranteed minimum output $y(t)$.

 ISBN: 978-1-61208-398-8

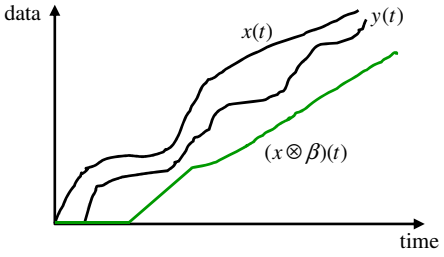


Figure 2. Convolution as a lower output bound

Definition 3 (Service Curve): Let system S with input flow $x(t)$ and output flow $y(t)$ be given. The system has a (minimum) service curve $\beta(t)$, iff $\beta(t)$ is a non-negative, non-decreasing function with $\beta(0) = 0$ and

$$y(t) \geq (x \otimes \beta)(t) \tag{5}$$

Figure 2 shows $(x \otimes \beta)(t) = \inf_{0 \leq s \leq t}\{x(s) + \beta(t-s)\}$ as lower bound of output $y(t)$ w.r.t. input $x(t)$.

Example 2: The commonly used *rate-latency function* reflects a service element which offers a minimum service of rate R after a worst-case latency of T, by doing so, all internal behavior of the service is hidden and only the worst-case is described.

$$\beta(t) = \beta_{R,T}(t) = R \cdot [t-T]^+ := R \cdot \max\{0; t-T\} \tag{6}$$

In Figure 4, the graph of $\beta_{R,T}(t)$ (green) depicts the rate-latency service curve with rate R and latency T.

Consider a system S with input flow $x(t)$, arrival curve $\alpha(t)$, output flow $y(t)$ and service curve $\beta(t)$. Then, according to [4], the following three bounds can be calculated.

Proposition 1 (Backlog, Delay and Output bound):

- Backlog bound v:
  $v(t) = x(t) - y(t) \leq \sup_{s \geq 0}\{\alpha(s) - \beta(s)\}$
- Delay bound $d(t)$ of input x (FIFO service):
  $d \leq \sup_{s \geq 0}\{\inf\{\tau : \alpha(s) \leq \beta(s+\tau)\}\}$
- Output bound $\alpha^*(t)$:
  $\alpha^*(t) = \alpha \oslash \beta := \sup_{s \geq 0}\{\alpha(t+s) - \beta(s)\}$

Figure 3 depicts d and v for a general arrival- and service curve.

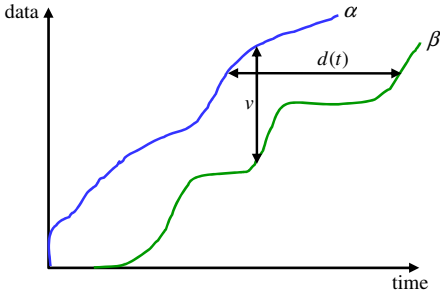


Figure 3. Backlog and delay bound

Example 3: Suppose a system with token bucket input and rate-latency service, according to (1), (3), (5), and based on

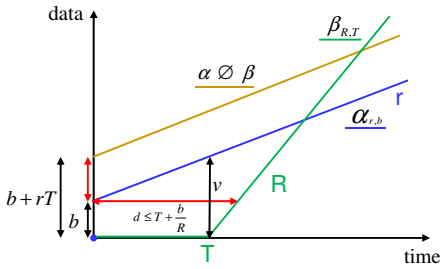


Figure 4. Example for all bounds, cf. Proposition 1

Proposition 1 the delay bound is given by $d \leq \frac{b}{R} + T$, the output bound by $\alpha^*(t) = r(t+T) + b$, and the backlog is bounded by $v = b + rT$ (see Figure 4).

III. Aggregate Scheduling

Up to now, only single flow-based scheduling has been considered. But, in many real-world systems *aggregate scheduling* is used [5]. We speak of aggregate scheduling whenever at least two streams are handled as a single stream. While in [6], delay bounds for FIFO networks are given, in [7] the main goal is to derive end-to-end delay bounds for general multiplexing. Important applications of aggregate scheduling are, among others, *differentiated service* domains (DS) of the Internet, the determination of safety-critical delay bounds within automotive intra car-communication [8], and diverse time-critical industrial applications [9].

In order to apply NC to such communication networks, we have to expand the rules to multiplexing and aggregate scheduling.

Assume that n flows enter a system (network) or system node which are scheduled by aggregation. According to [10] the aggregate input flow and arrival curve are given as follows.

Definition 4 (Aggregation, Multiplexing): An *aggregation* (also: *multiplexing*) of n flows can be expressed by adding the single input flows and arrival curves, respectively:
let $n = 2$, then the aggregated input flow is $x(t) = x_1(t) + x_2(t)$ and $\alpha(t) = \alpha_1(t) + \alpha_2(t)$, where x_1, x_2 and α_1, α_2 are the corresponding single input flows and arrival curves.

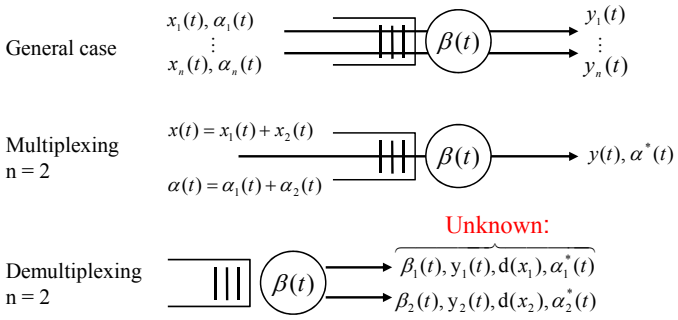


Figure 5. Multiplexing of flows: input x_i, output y_i, arrival & service curve α_i, $\beta(t) = \beta_{aggr}$

Considering multiplexed streams (Figure 5) the question is whether Proposition 1 might be applied to the individual

streams of an aggregation, for example to calculate the maximum delay of a single flow x_i.

Firstly, this depends on the type of aggregate scheduling, like FIFO (see [11]), priority-scheduling, or unknown arbitration, and secondly on the service curve β_{aggr} of the aggregated flow. If no knowledge about the choice of service between the flows is present, then we speak of *arbitrary multiplexing* [12] (also: *blind multiplexing*), and the situation is more complex. In such cases, the distinction between *strict* and *non-strict* aggregate service curves plays an important role [4].

Proposition 2 (Blind Multiplexing): Let a node be serving the flows x_1 and x_2 with an unknown arbitration. Assume that the node guarantees a strict service curve β to the aggregation of the two flows, and that flow x_2 is bounded by α_2. Let $\beta_1(t) := [\beta(t) - \alpha_2(t)]^+$; β_1 is a service curve for flow x_1 if it is wide-sense increasing.

Definition 5 (Strictness of service curve): A system S offers a *strict* service curve β to a flow, if during any backlogged period $[s,t]$ of duration $t-s$ the output y of the flow is at least equal to $\beta(u)$, i.e., $y(t) - y(s) \geq \beta(t-s)$. Obviously, any strict service curve is also a service curve.

Example 4: Figure 6 (left) shows a token bucket input $x = rt + b$ and a service curve $\beta(t) = R \cdot t$. Here, the output $y(u) \geq \beta(u)$ in all backlogged periods u: $u \leq$ busy period. Thus, β is *strict* (there is only one backlogged period u).

If we switch to the rate-latency service curve $\beta_{R,T}(t) = R \cdot [t-T]^+$ (Figure 6, right), we get a *non-strict* service curve. The backlogged period starts at 0 and never ends: because in the worst case, all input data of x remains in the system for time T before being served with rate R, but new data of x always arrives during T. The definition of the service curve specifies the output y as $y(t) \geq (x \otimes \beta)(t)$. Indeed, it is valid that the output $y(u_0) \geq \beta_{R,T}(u_0)$, but this is not guaranteed regarding the backlogged period $u > u_0$. Thus, it is possible that $y(u) \ngeq \beta_{R,T}(u)$ as $(x \otimes \beta_{R,T})(u) - (x \otimes \beta_{R,T})(0) = (x \otimes \beta_{R,T})(u) < \beta_{R,T}(u) = \beta_{R,T}(u) - \beta_{R,T}(0)$ if $T > 0$. In this scenario, β is *non-strict*.

The above example raises the issue whether there are classes of service functions that are always *strict* or *non-strict*, respectively.

In literature, the service curve $\beta_{R,T}(t) = R \cdot [t-T]^+$ (or even any *convex* service curve) is often used as a strict service curve per se; see, for instance, [13]. We will see that both, *strictness* and *non-strictness*, are not only based on the service curve itself but also on the corresponding input flow x. Considering aggregate flow situations, this means that the strictness of an aggregated input flow needs to be checked before applying the important Proposition 2. Therefore, the condition $y(u) \geq \beta(u)$, $\forall$ backlogged periods u needs to be proofed.

First, we will provide and prove some characterizations for applications using token bucket input flows and the commonly used rate-latency service curves.

Theorem 1 (Non-strict functions): Let a system with rate-latency service curve $\beta_{R,T}$ and token bucket arrival curve $\alpha_{r,b}$ with $r < R$ and $T > 0$ be given. Furthermore, the worst case scenario is assumed: the input is served with minimum rate R after a possible maximum delay T. We claim that the service curve $\beta_{R,T}$ cannot be strict, if the input flow $x(t)$ is a *strictly increasing* function.

Proof:
Assume $\beta_{R,T}$ is strict. Based on Proposition 1 we know $\alpha^*(t) = \alpha \oslash \beta := \sup_{s \geq 0}\{\alpha(t+s) - \beta(s)\}$, here $\alpha^*(t) = r(t+T) + b$. Because $r < R$, there is a point in time t_s, so that $\beta_{R,T}(t_s) = \alpha^*(t_s)$ and $\beta_{R,T}(t) > \alpha^*(t)$ if $t > t_s$, i.e., $\forall\ t_0 > t_s : \beta_{R,T}(t_0) - \alpha^*(t_s) \geq \alpha^*(t_0) - \alpha^*(t_s)$ $\Rightarrow \Delta\beta_{R,T} = \beta_{R,T}(t_0) - \beta_{R,T}(t_s) \geq \alpha^*(t_0) - \alpha^*(t_s) = \Delta\alpha^*$. Since $x(t)$ is strictly increasing and latency $T > 0$: for any $t_0 > t_s$: $u := t_0 - t_s$ is a backlogged period—cf. Figure 6 (right).
$\beta_{R,T}$ is supposed to be strict, so output $y(u) \geq \beta_{R,T}(u) = \beta_{R,T}(t_0) - \beta_{R,T}(t_s) \geq \alpha^*(t_0) - \alpha^*(t_s) = \alpha^*(t_0 - t_s)$. But this is a contradiction to α^* being an arrival curve for output y. Thus, the assumption is wrong, i.e., $\beta_{R,T}$ is non-strict. □

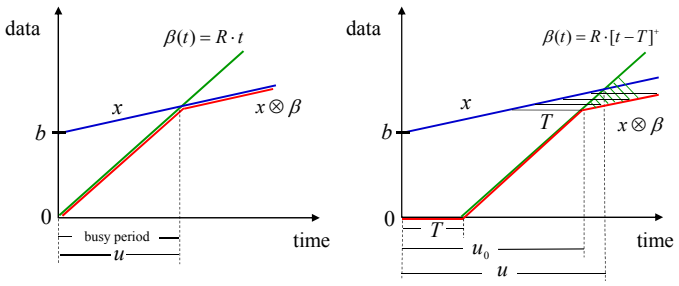


Figure 6. Strict and non-strict service curve

Unfortunately, the feature of being a non-strictly increasing input x is not a sufficient condition for a strict service curve $\beta_{R,T}$: using the same token bucket arrival curve $\alpha_{r,b}$ and rate-latency service curve $\beta_{R,T}$, one can find non-strictly increasing input functions x that make the service curve $\beta_{R,T}$ either strict or non-strict. This will be demonstrated by the following example.

Example 5: Let

$$\alpha_{r,b} := \begin{cases} 1,5t + 5 & \text{for } t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and $\beta_{R,T} := 2(t-2)^+$. Furthermore, let the input x be first identical to $\alpha_{r,b}$, and then stagnate at time t'. The parameter t' is computed using equation $\alpha_{r,b}(t) = \beta_{R,T}(t+T)$. This guarantees that no displacement of the $\beta_{R,T}$ graph within the convolution graph of $x \otimes \beta_{R,T}$ occurs:
$1,5t + 5 = 2((t+2) - 2); t = t' = 10$ fulfills this equation. So, we define the input as

$$x(t) = \begin{cases} 0 & \text{for } t \leq 0 \\ 1,5t + 5 & \text{for } t \leq 10 \\ 20 & \text{otherwise} \end{cases} \tag{8}$$

Result: The service curve $\beta_{R,T}$ is strict (Figure 7).

Now, the input x is changed from x to $\tilde{x}$ ($\tilde{x}$ is still non-strictly increasing):

$$\tilde{x}(t) = \begin{cases} 0 & \text{for } t \leq 0 \\ 0,75t + 2,5 & \text{for } t \leq 10 \\ 10 & \text{otherwise} \end{cases} \quad (9)$$

Result: The same service curve $\beta_{R,T}$ is now non-strict (Figure 7).

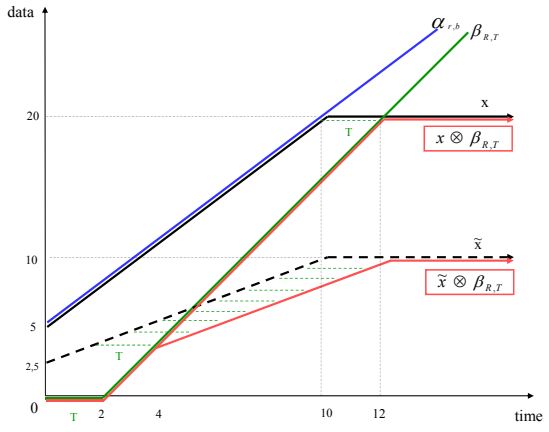


Figure 7. Input x changed to $\tilde{x}$ causes non-strictness

Obviously, all **input functions *x*** of the form (or multiple repetitions)

$$x(t) = \begin{cases} mt + n & \text{for } t \leq t' \\ const & \text{otherwise} \end{cases} \quad (10)$$

cause the service curve $\beta_{R,T}$ to be strict, if the constant part of x starts within or on the border of the red-dashed triangle of Figure 8.

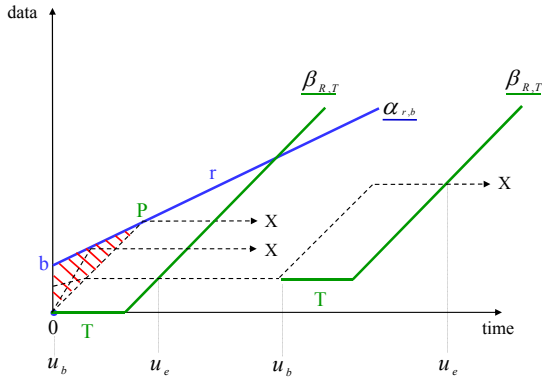


Figure 8. Input area making $\beta_{R,T}$ strict

Here, u_b is the begin and u_e the end of a backlogged period, b is the burst size of the arrival curve $\alpha_{r,b}$ and $P = P(\acute{t}, \acute{y})$ with $\acute{t}$: $\alpha_{r,b}(\acute{t}) = \beta_{R,T}(\acute{t} + T)$, i.e., the intersection of $\alpha_{r,b}$ with the parallel line to $\beta_{R,T}$, given by the curve $y = Rt$.

So, we state: the feature of being *strict* or *non-strict* is not only determined by a service curve itself. In all cases, this feature depends on both, the service curve *and* the actual input x. Since often concrete input flows remain unknown, this aggravates NC's application to worst-case analysis of scenarios including aggregate scheduling.

IV. Bypassing the Strict-Non-Strict Problem

In the following, we try to overcome the problem of strictness or non-strictness in case of aggregate scheduling. Our aim is to look for and—if possible—create a service curve for a single flow x_i of an aggregate flow x without any consideration, either on FIFO or non-FIFO scheduling, and strictness (non-strictness) of the aggregate-service curve $\beta_{aggr}(t)$.

All together, these considerations provoke the following statement that determines a generalization of Proposition 2 as it does not differentiate between strict and non-strict:

Theorem 2 (Construction of always-to-use service curves): Consider a node with some unknown arbitration serving the flows x_1 and x_2. Let x be the aggregated input with $x = x_1 + x_2$ and $y = y_1 + y_2$ the aggregated output, respectively. Assume that the node offers a service curve β to the aggregation of the two flows with $y(t) \geq (x \otimes \beta)(t)$, and flow x_2 is α_2-smooth. Define $\beta_1(t) := [(x \otimes \beta)(t) - \alpha_2(t)]^+$. If β_1 is wide-sense increasing, then β_1 is a service curve for flow x_1.

Proof:

Since $\beta = \beta_{aggr}$ is a service curve for input x, we know:

(i) $y \geq x \otimes \beta_{aggr}$

Let $\Phi := \{f \mid f(t_1) \geq f(t_0)$ for $t_1 \geq t_0,\ f(t) = 0$ for $t < 0,\ f(t)$ left-continuous, $t \in \mathbb{R}\}$

In [4] we find the following algebraic rule:

If $f(0) = g(0) = 0$ then $(f \otimes g \leq min\{f, g\} \Rightarrow$

$x \otimes [(x \otimes \beta_{aggr})] \leq min\{x, (x \otimes \beta_{aggr})\} \Rightarrow$

$x \otimes [(x \otimes \beta_{aggr})] \leq x \otimes \beta_{aggr} \leq y.$

Thus, $x \otimes [(x \otimes \beta_{aggr})] \leq y \Rightarrow$ expression $x \otimes \beta_{aggr}$ itself is a service curve for x, and together with (i) it is even *strict*. But this means: $\beta_1(t) := [(x \otimes \beta)(t) - \alpha_2(t)]^+$ is a service curve for flow x_1 if β_1 is wide-sense increasing. □

Based on Theorem 2, we can derive the construction of a strict service curve.

Theorem 3 (Construction of a strict service): Suppose a node with the conditions of Theorem 2 and with a service curve β to the aggregate with $y(t) = (x \otimes \beta)(t)$ instead of $y(t) \geq (x \otimes \beta)(t)$.

Then the following is always valid:

(i) $\beta^* := (x \otimes \beta)$ is a strict service curve to the flow x

(ii) $\beta^* := (x \otimes \beta)$ is the greatest strict service curve to x

Proof:

(i): β is service curve, i.e., $y \geq x \otimes \beta \ \Rightarrow\ y(t) \geq \beta^*(t)$ which means β^* is strict.

(ii): Assume β^* is not the greatest service curve, that means there is a β' with $\beta \geq \beta' > \beta^*$ and β' is strict. Then $y \geq \beta'$, and $\Rightarrow\ y = x \otimes \beta \geq \beta'$. Since $\beta^* := (x \otimes \beta) \Rightarrow \beta^* \geq \beta'$ which contradicts the assumption. □

Thus, we can state the following: the construction of a strict service curve is possible in case of blind multiplexing with non-strict service curve to the aggregate input. Of course, if $y(t) \geq (x \otimes \beta)(t)$ and not $y(t) = (x \otimes \beta)(t)$, then it may be that a greater strict service curve than $\beta^* := (x \otimes \beta)$ exists

 ISBN: 978-1-61208-398-8

(and therefore a better one w.r.t. β_i of a single flow x_i); but in any case, it is a strict service curve.

Now we give another theorem to overcome the question of strictness or non-strictness of a service curve. Due to Theorem 1, the important service curve $\beta_{R,T}$ cannot be strict, if input flow $x(t)$ is a strictly increasing function. And, according to Figure 7, the change of input x to $\tilde{x}$ causes the service curve $\beta_{R,T}$ to be non-strict although $\beta_{R,T}$ is not changed. Consequently, the assumption of Proposition 2 would not be fulfilled—but sometimes it is possible to characterize an (aggregated) input $x = x_1 + x_2$ so that a service curve $\beta_1(t)$ of flow x_1 exists, even if the service curve β of the aggregate is non-strict:

Theorem 4 (Singular flow bounded by K): Consider a node serving the flows x_1 and x_2 with some unknown arbitration between the two flows. Let x be the aggregated input with $x = x_1 + x_2$ and $y = y_1 + y_2$ the aggregated output, respectively. Assume that the node offers a service curve β to the aggregate of the two flows, and let flow x_2 be bounded by $K > 0$. Define $\beta_1(t) := [\beta(t)] - K]^+$. If β_1 is wide-sense increasing, then β_1 is a service curve for flow x_1.

Proof:

Let $\Phi := \{f \mid f(t_1) \geq f(t_0)$ for $t_1 \geq t_0,\ f(t) = 0$ for $t < 0,\ f(t)$ left-continuous, $t \in \mathbb{R}\}$

According to the algebraic rules in [4]:

Rule a): $K + (g \otimes f) = g \otimes (K + f)$ if $g, f \in \Phi,\ K \in \mathbb{R}^+$

Rule b): If $f \leq f^{'}$ and $g \leq g^{'} \Rightarrow f \otimes g \leq f^{'} \otimes g^{'}$ for $g, g^{'}, f, f^{'} \in \Phi$, β is service curve, i.e., $[y_1(t) + y_2(t)] \geq ((x_1 + x_2) \otimes \beta)(t)$

$\Rightarrow y_1(t) \geq ((x_1 + x_2) \otimes \beta)(t) - y_2(t) \Rightarrow$ Because $x_1 + x_2 \geq x_1$ and Rule b):

$\Rightarrow \ y_1(t) \geq (x_1 \otimes \beta)(t) - y_2(t)$. Always is true $y_2(t) \leq x_2(t)\ \forall t \ \Rightarrow y_1(t) \geq (x_1 \otimes \beta)(t) - x_2(t)$, and since by assumption $x_2(t) \leq K \Rightarrow y_1(t) \geq (x_1 \otimes \beta)(t) - K$.

Applying Rule a) we get $(x_1 \otimes \beta)(t) - K = (x_1 \otimes (\beta - K)(t)$ $\Rightarrow y_1(t) \geq (x_1 \otimes (\beta - K)(t)$; which means $\beta_1 := [\beta - K]^+$ is service curve of x_1 if it is wide-sense increasing.□

Remark:

Now, we can state that all (aggregate-) input functions $x = x_1 + x_2$ with $x_2(t) \leq K$ possess the service curve $\beta_1 := [\beta - K]^+$ independent from strictness or non-strictness of β, as for example of the aggregate x:

$$x := \begin{cases} mt + n & \text{for any } t \in \mathbb{R}^+ \\ const & \text{otherwise} \end{cases} \tag{11}$$

Example 6: A typical scenario of blind multiplexing is given by two input flows x_{low} and x_{high} with a worst case service situation for flow x_{low}—also called *preemptive priority schedule*, i.e., x_{high} will be served first, and will always interrupt the x_{low}-service as soon as data of flow x_{high} is present. Only if there is no data of flow x_{high}, flow x_{low} will be served.

Proposition 2, Theorem 2, and Theorem 4 may provide a service curve β_{low} for the low-priority flow x_{low}. An important question is whether this (minimum) service curve β_{low} together with α_{low}— if known—can be used to compute the maximal delay, the maximal backlog bound, or the maximal output bound of the flow x_{low} by applying Proposition 1. In terms of backlog bound we can argue that the service facility is work-conserving, i.e., the 'unfinished work' $x(t) - y(t)$ depends only on the arrival instants and the data(packet-)length, and not on the order of service. Therefore, it is easy to realize that Proposition 1 (Backlog bound) is applicable here:

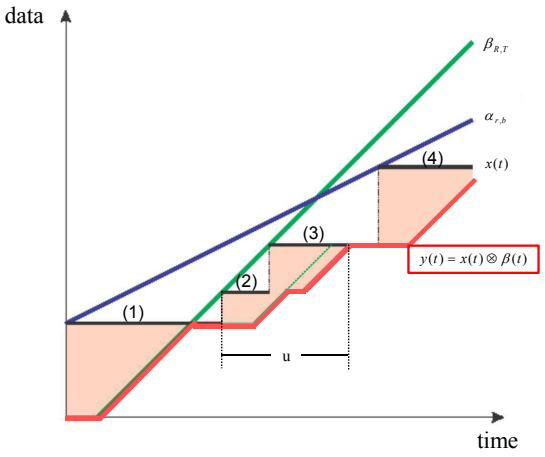


Figure 9. Finite backlogged period u—but $\beta_{R,T}$ is not strict

$$x_{low}(t) - y_{low}(t) \leq \sup_{s \geq 0}\{\alpha_{low}(s) - \beta_{low}(s)\} \tag{12}$$

Nevertheless, relating to x_{low}-traffic, both of the other bounds, i.e., Delay and Output bound—are applicable, too. For example, considering the delay bound of x_{low}:

$$d \leq \sup_{t \geq 0}\{\inf\{\tau : \alpha_{low}(t) \leq \beta_{low}(t + \tau)\}\} \tag{13}$$

Here, the *virtual delay*

$$d_\tau(t) = \inf\{\tau \geq 0 : \alpha_{low}(t) \leq \beta_{low}(t + \tau)\} \tag{14}$$

ensures that an input x which arrives at time t will leave service not later than $d_\tau(t)$. This is guaranteed for FIFO scheduling but not for blind multiplexing. However, we may presume FIFO per single flow (e.g. x_{low} within the aggregate of blind multiplexing) and thus apply all three bounding theorems without any restrictions.

The important statement $d \leq \sup_{s \geq 0}\{\inf\{\tau : \alpha(s) \leq \beta(s + \tau)\}\}$ is only valid for FIFO systems. Hence the question is whether it is possible to state a similar proposition for systems in general—either FIFO or not. The next theorem is a first answer to this.

Theorem 5 (Delay bounds in general): Let a system S with input x, arrival curve α, output y, service curve β, and point in time t_α with $\forall t > t_\alpha : \alpha(t) < \beta(t)$ be given. Furthermore, let $U = \{u | u$ is backlogged period$\}$, and $l(u)$ be the length of a backlogged period. If the service curve β is strict or u is finite $\forall u \in U$, then the maximal delay d is given by

$$d \leq \sup_{u \in U}\{l(u) : (x \otimes \beta)(t) \leq x(t) \wedge t \in u\} \tag{15}$$

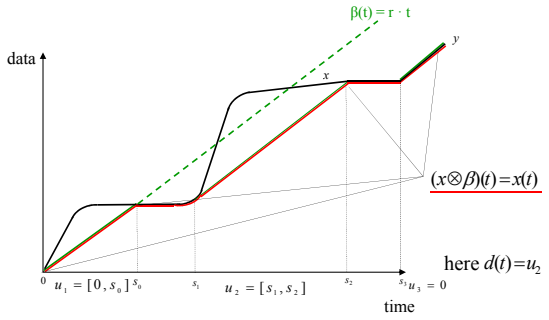


Figure 10. Strict service curve

Proof:
Case: u is finite $\forall u \in U$—proof is trivial.
Case: service curve β is strict: due to the subsequent Lemma the backlogged period is finite. And otherwise a backlogged period is finite at time t if $x(t) = y(t)$, cf. Figure 10. Take any backlogged period u and let s be start and t be the end of the backlogged period $u = [s, t]$, that means for input x and output y: $y(s) = x(s)$ and $y(t) = x(t)$. Let β be the service curve, i.e., output y is lower-bounded by $x \otimes \beta$; now, $\forall$ subsets $v \subseteq u$ with $v := [s, z]$ and $z \leq t$:
(i) $x(z) \geq y(z) \geq (x \otimes \beta)(z)$.
On the other hand: $\forall\ \Delta_v := z - s$, and due to strictness of β: $y(\Delta_v) \geq \beta(\Delta_v) = \beta(z-s)$. That means we get the following equation:
(ii) $\inf_{s \leq \tilde{t} \leq z}\{x(s)+\beta(\tilde{t}-s)\} = x(s)+\beta(\tilde{t}-s) = (x \otimes \beta)(\tilde{t})$.
This is because at the points from s to z inside the finite backlogged period $u = [s, t]$, the value of convolution $x \otimes \beta$ is determined by the service curve β alone. Therefore, for $z \to t$ we get:
(iii) $y(z) = y(t) = (x \otimes \beta)(t)$ and $y(t) = x(t) \ \Rightarrow x(t) = (x \otimes \beta)(t)$ where t is the end of backlogged period u.
$\Rightarrow l(v) \leq l(u)$ with $(x \otimes \beta)(z) \leq x(z)$.
$\Rightarrow$ Maximal delay d: $d \leq \sup_{u \in U}\{l(u) : (x \otimes \beta)(t) \leq x(t) \wedge t \in u\}$. $\square$

Lemma: Let a system S with input x, arrival curve α, output y, service curve β, and point in time t_α with $\forall t > t_\alpha$: $\alpha(t) < \beta(t)$ be given. If the service curve β is strict, then any backlogged period is finite.

Proof:
Since β is strict we have $y(t) \geq \beta(t)\ \forall t$.
Of course $y(t) \leq x(t)$ is always true, altogether

$$(x \otimes \beta)(t) \leq \beta(t) \leq y(t) \leq x(t) \quad \forall t \tag{16}$$

Let u be any backlogged period, and suppose u is *not* finite. Then $y(t) < x(t)\ \forall t$ and, therefore, also $\forall t > t_\alpha$.
From (16): $(x \otimes \beta)(t) \leq \beta(t) \leq y(t) < x(t) \leq \alpha(t-0) = \alpha(t)$. But that means $\beta(t) < \alpha(t)\ \forall t > t_\alpha$ which contradicts the precondition of the Lemma. Thus, u is finite, i.e., $\exists\ t_0$: $y(t_0) = x(t_0)$. $\square$

Unfortunately, the opposite statement is not valid as shown in Figure 9.

V. Conclusions

This paper deals with worst case modeling of aggregate scheduling. We want to get guaranteed performance parameters, like maximal end-to-end delay of individual so-called left-over flows of an aggregate. When using the analytical tool Network Calculus (NC), among others the service curve is required as main modeling element. In case of blind multiplexing the following particular problem occurs: the construction of a service curve for the single output after demultiplexing an aggregated flow $x = x_1 + x_2$ requires the *strictness* of the aggregated service curve.

In publications like [13] [12], or others, it is assumed that the rate latency service curve $\beta_{R,T}$ (very often used as aggregated service curve) fulfills the strictness property.

In this article, we demonstrated that the property of being *strict* or *non-strict* does not depend on the service curve solely. Only in combination with the concrete input—or at least with a special class of input—one can decide whether a service curve is strict or non-strict.

By providing and proving, firstly, theorems to get weaker forms of strictness and, secondly, a more general approach to get service curves or worst case delay bounds, we bypassed this difficulty. Therefore, deterministic performance analysis based on NC in situations comprising aggregate scheduling remains applicable.

References

[1] U. Klehmet, T. Herpel, K.-S. Hielscher, and R. German, "Worst Case Analysis for Multiple Priorities in Bitwise Arbitration," in *GI/ITG-Workshop MMBnet 2007, Hamburg, pp. 27-35*.
[2] J. Liebeherr, A. Burchard, and F. Ciucu, "Delay Bounds in Communication Networks with Heavy-tailed and Self-similar Traffic," *IEEE Trans. Inform. Theory*, vol. 58(2), pp. 1010–1024, 2012.
[3] R. Cruz, "A calculus for network delay, part i: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37-1, pp. 114–131, 1991.
[4] J.-Y. Le Boudec and P. Thiran, *Network Calculus*. Springer Verlag LNCS 2050, 2012.
[5] Y. Ying, F. Guillemin, R. Mazumdar, and C. Rosenberg, "Buffer Overflow Asymptotics for Multiplexed Regulated Traffic," *Performance Evaluation*, vol. 65-8, 2008.
[6] A. Charny and J.-Y. Le Boudec, *Delay Bounds in a Network with Aggregate Scheduling*. Springer Verlag LNCS 1922, 2000.
[7] J. Schmitt, F. Zdarsky, and M. Fidler, "Delay Bounds under Arbitrary Multiplexing," *Technical Report*, vol. 360/07, 2007.
[8] T. Herpel, K.-S. Hielscher, U. Klehmet, and R. German, "Stochastic and Deterministic Performance Evaluation of Automotive CAN Communication," *Computer Networks*, vol. 53, pp. 1171–1185, 2009.
[9] S. Kerschbaum, K.-S. Hielscher, U. Klehmet, and R. German, "A Framework for Establishing Performance Guarantees in Industrial Automation Networks," in *Proceedings MMB and DFT 2014, Bamberg, pp. 177-191*, March 2014.
[10] M. Fidler and V. Sander, "A Parameter based Admission Control for Differentiated Services Networks," *Computer Networks*, vol. 44, pp. 463–479, 2004.
[11] G. Rizzo, "Stability and Bounds in Aggregate Scheduling Networks," Ph.D. dissertation, Ecole Polytechnique Federale De Lausanne, 2008.
[12] J. Schmitt, F. Zdarsky, and I. Martinovic, "Improving Performance Bounds in Feed-Forward Networks by Paying Multiplexing Only Once," in *Measurements, Modelling and Evaluation of Computer and Communication Systems(14th GI/ITG Conference), Dortmund*, March 2008.
[13] A. Bouillard, B. Gaujal, and S. Lagrange, "Optimal Routing for End-to-end Guarantees: the Price of Multiplexing," in *Valuetools '07, Nantes*, October 2007.

 ISBN: 978-1-61208-398-8

Extending Contemporary Network Modeling Towards the Photonic Layer

Jan Kundrát and Stanislav Šíma

CESNET, z.s.p.o.,
Praha, Czech Republic
E-Mail: jan.kundrat@cesnet.cz

***Abstract*—A significant interest recently sprouted towards describing a production computer network in a machine-readable way. Most of the effort has, however, so far focused on providing an accurate description starting at the L2/L3 of the ISO/OSI model. Such an approach cannot accommodate advanced applications, such as accurate transfer of ultra-stable frequency signal. In this work, we aim at outlining possible ways of extending the existing network models towards describing the photonic properties of the contemporary networks. When a network model allows for an accurate description of the photonic substrate, it becomes suitable for automated verification of critical properties which are required by emerging network users.**

Keywords–photonics; optical networking; network modeling; CAD.

I. Introduction

When dealing with contemporary networks, a list of interesting properties includes much more than just the available bandwidth and an administrative cost of a given interconnect. Optical networking, or photonic networking, is a novel way of designing networks where the light which travels through a fibre link is not converted to electricity at various points [1]. A photonic network is capable of carrying concurrent optical signals between the connected nodes. Properties which are sufficient for a reasonably accurate description of a generic packet-switched network no longer make sense for a photonic service. It is irrelevant whether the provisioned transceivers at exchange points support 100 Gbps Ethernet. What matters is how the light is affected by the fibre paths, or by the active devices which amplify, switch and otherwise modify the carried signal.

Having a well-established model of how a photonic network (or, indeed, any computer network in general) is constructed allows for many advanced applications. If the model includes physical properties like location of the actual hardware devices, it can be used for optimizing dispatching the technicians in case of an outage. A well-designed abstraction scheme along with a well-maintained inventory can be used for checking the bill-of-materials needed when building networks. Model checking and verification can be used to verify upfront that a network which is about to be built, with fibre buried a few meters deep, is capable of achieving the properties its owners expect, and are willing to pay for.

This paper starts by explaining the state of art in the area of network modeling in Section II. In Section III, we explain what properties of the photonic substrate have to be tracked in the model, and outline how to cover them as an extension of an existing modeling language. The article wraps up with a plan for future work and conclusion in Section IV.

II. Related Work on Network Modeling

Different use cases require varying level of detail to be present in a model of the network. As an example, a typical network router usually needs to know where to forward packets destined for a particular IP address. On the other hand, a Network Operating Center (an *NOC*) usually has more stringent requirements. In our use case we are going to focus on modeling the complete topology of a network, i.e., a situation where at least one node is aware of every property of the network being modelled. This is in contrast to a distributed system where each node typically maintains just a subset of the required information.

In this work, we are investigating universal models, which are capable of conveying a wide range of information. We are looking for a single model which is usable for (or at least extensible to) describing the actual computer network on many levels.

A. Network Description Language

The Network Description Language (NDL) is a format which was introduced by Jeroen van der Ham [2]. Based on the RDF (Resource Description Framework) standard [3], the NDL builds a distributed set of documents, each referenced by a special URL (Uniform Resource Locator), along with relations among the described entities. The basic NDL contains three classes of modeled entities:

Location which describes a physical location where *Devices* are to be found in real-world.

Device which represents a physical piece of equipment which is a part of the network.

Interface which is used to model an interface or port of the physical *Device* which is used for connecting to other devices.

These objects or classes are connected by various predefined relations. The NDL format predefines these six properties:

locatedAt for tying together *Locations* and *Devices*.

hasInterface for assigning *Interfaces* to *Devices*.

connectedTo which represents an external connection of two *Interfaces*.

switchedTo which, compared to the `connectedTo`, represents an internal connection, presumably within a single *Device*.

name for conveying the name of a resource in a RDF way.

description which includes a free-format, human-readable description to any class.

The original, baseline NDL did not offer much functionality. While its features were sufficient for describing the *topology* of the network, no provisions were included for adding machine-readable properties such as the type of actual link, or available capacity. Due to the nature of the RDF format, though, the NDL itself was very extensible. As a logical next step, its main authors built on top of the existing standard and attempted to introduce the NDL concept to contemporary optical networks in their follow-up work [4]. It should be noted that while the article in which this extension was first presented was titled *Using the Network Description*

 ISBN: 978-1-61208-398-8

Language in Optical Networks the word *"optical"* does not refer to the optical properties of photonic networks.

This refinement of the NDL language added (among other changes) a single new class, a **Link**, which represents a previously implied connection between two entities, and the following new properties:

capacity which provides the provisioned throughput over a *Link*.

encodingType which defines the encoding used on a given *Link*.

encodingLabel which extends the information provided by `encodingType`.

These properties addressed one of the major shortcomings of the original NDL specification, as useful properties were now assignable to individual links. A perfect example is the `encodingType`, which enabled specifying the type of a link, such as whether it represents an L2 Ethernet segment, or a lower level lambda path. There was still little to no support for machine-readable analysis of the attached information, though, and the model did not consider multiple layers which constitute modern computer networks.

As shown by Dijkstra [5], it is rarely sufficient to describe just a single layer of the network at a time. Similarly, it is often not feasible to always operate on the full node graph of a moderately-sized network. The NDL format was therefore extended by provisions for multi-layer network modeling, as demonstrated by van der Ham [6].

The authors show how the NDL can be extended with a concept of *adaptation* and *multiplexing*. Together, these features allow modeling common concepts such as aggregating multiple VLANs into one Ethernet port. The concept is further extended with support for the so-called *Switch Matrix*, a construct suitable for describing features which the network provides on each of the represented layers.

It should be noted that, as Dijkstra shows, opening up the model for multilayer description brings along unintuitive situations. As an example, he presents a use case [7] where the only feasible path within a laboratory in *Universiteit von Amsterdam* and the *Université du Quebec* actually crosses the same fibre *twice*.

An attempt at extending the NDL for photonic services was conducted at CESNET. In a deliverable of the Phosphorus project, CESNET contributed an add-on schema [8] suitable for conveying a wide range of properties concerning the optical path. The proposed extension was capable of accurately describing specialized equipment, such as the DCM (Dispersion Compensation Module, a device which corrects for phenomena resulting from different propagation speeds of different wavelengths in the fibre medium, including its effect on a single-color modulated light) as well as actual, real-world fibre spans and optical light paths deployed in the CESNET2 network at that time. However, the proposal did not make use of the usual NDL features such as the concept of adaptations, and it was not formally merged into the upcoming revisions of the NDL standard.

B. Network Markup Language and the Network Service Interface

The NDL language described in the previous section was not the only contender for delivering a format capable of accurately describing multilayer computer networks. The GÉANT collaboration worked on the cNIS [9], while the US-based ESnet was independently working on a similar solution needed by perfSONAR. The Network Markup Language (NML) working group [10] was eventually formed within the Open Grid Forum [11] to create a unified, multi-layer network model.

Roughly in parallel, an effort was ongoing on delivering a unified specification for communicating between networks of different administrative contexts. There was a huge range of envisioned applications, from bandwidth-hungry users of the Bandwidth-on-Demand (BoD) service, to testbed users and network students to experiment on a real network, yet in a controlled and harmless manner. These applications, however, required understanding of the underlying network topology.

The Network Service Interface (NSI) framework [12] and its associate suite of protocols is designed to facilitate establishing of multi-domain network connections using a RESTful (Representational State Transfer) web service API (Application Programming Interface). As of 2014, the NSI and the associated specifications are subject of active research and development work. There are plans to use the resulting software throughout the newly developed Testbed as a Service (TaaS) offering by GÉANT. The network topologies represented by NML are being studied for further uses, such as multi-constrained path selection [13], a process where the choices of a link path are affected by several parameters. It seems that the academy has finally settled on a single modeling framework after a phase where concurrent projects were each pushing for their own solution.

C. Nodes, Links and Adaptations

A common feature found in each sufficiently mature multi-layer or multi-domain schema is the appearance of the following building blocks:

Nodes for entities which represent active network devices, or abstraction points which serve a certain purpose which is visible to the other layers.

Links or *Connections* whose purpose is to represent interconnections between the *Nodes*. Each link typically carries a set of *Properties*.

Adaptations which are capable of interconnecting *Nodes* belonging to conceptually different layers of the abstraction.

Within the NDL, these concepts were accurately represented by the *Device*/*Interface* and *Link* entities. Under the NML abstraction, these are realized through the *Node*/*Port*, *Link* and *Adaptation*/*SwitchingMatrix* classes. On a basic level, these building blocks are enough for representing the contemporary network hardware as long as a mechanism for tagging the entries with appropriate machine-readable properties is supported.

At CESNET, the author is developing a graphical CAD (Computer-Aided Design) application which should be capable of handling the future development of the NML scheme as long as the additions and changes remain within the bounds of these three broad categories of functions. Internally, the application uses these classes for manipulating a hierarchy of objects which correspond to actual real-word entities, such as optical amplifiers, Ethernet switches or a leased line between two points of presence (PoP) on a network. The clean architecture of the NML specification enables its users to be written in a generic way. A compliant application will require no code changes to support an updated or new NML schema. The

 ISBN: 978-1-61208-398-8

ultimate goal is to have all the required information present in the newly developed schema; that is, the computer program, which is the user of the extended NML, shall be able to infer any required details from the schema itself.

III. Accommodating the Photonic Layer

While the NSI suite and the associated NML protocol are generic, support for describing the physical layer is explicitly said to be out of scope for the core specification [14]. However, based on personal contact with the GLIF (Global Lambda Integrated Facility, an international consortium that promotes lambda networking) members who are working on the NML adoption, the project is open to external contributions adding support for additional schemes.

A. Importance of the Photonic/Optical Properties

The contemporary Internet, and computer networks in general, are very good at delivering high bandwidth with reasonable latency and mid-to-high jitter. (Jitter describes how much a latency between the end points, which are communicating together fluctuates over time. A low jitter means that the latencies are stable, while high jitter indicates that the latencies might be very low at one point in time, but can grow significantly at any instant.) These properties are usually sufficient for a wide range of applications, from modest HTTP (Hypertext Transfer Protocol) requests and interactive SSH (Secure Shell) or RDP (Remote Desktop Protocol) traffic on one hand to VoIP (Voice over Internet Protocol) or media streaming applications at the other end of spectrum. It is usually acceptable that a latency of each packet varies between, e.g., 15 ms and 45 ms, and the applications which are more sensitive to increased jitter can usually cope with this phenomenon through additional buffering to smooth-over the variance.

However, there are other sorts of applications which are much more demanding, to an extent where a classic network equipment cannot keep up the pace [15]. A photonic service guarantees that the optical signal is transmitted throughout the network without any conversion between the light and electricity, or even statistical processing by a DSP (Digital Signal Processing) circuit common in contemporary systems. The signal is amplified along the fibre paths by dedicated, all-optical amplifiers, and circuit switching is implemented by all-optical equipment. The existing applications which require a photonic service include for example extremely accurate frequency transfers, which are used for comparison of national-level atomic clocks, precise instruments which are involved with the maintenance of the UTC (Coordinated Universal Time) time standard, or other metrological applications [16].

Being able to understand the underlying network opens up opportunities for more intelligent system procurement and significant cost savings [17, p. 20-22], both due to improved network design and innovative concepts such as bidirectional single fibre transmission [1].

B. Extending NML towards the Photonic Layer

Within a computer network which is capable of providing a photonic service, the emphasis on the tracked properties shifts down to the lower layers of the transmission stack. The total available bandwidth of a link is no longer the focus of the model; instead, the data describe how each link can transfer the light, and what modifications are inflicted on the signal as it travels through the fibre.

The total attenuation is one of these properties. However, as the total loss of signal is dependent on the frequency, the attenuation is not a single number, but instead a function, which computes the total loss based on the wavelength, or color of the light beam which passes through the fibre. Because the fibre usually conforms to one of predefined quality/performance classes, it makes sense to track the properties of each of the fibre type separately, and calculate the expected loss as a function of the fibre class and the total fibre length. However, in real world, the attenuation is not affected just by the fibre length. Mechanical stress or vibrations along the fibre path affect the transmission, too, and it is unfortunately common for construction works to interfere with a buried cable in a catastrophic mode, which typically involves an excavator's bucket. When the fibre is spliced back into a working state, the newly introduced joint deteriorates the overall transmission properties, and as such, the measured attenuation increases. Each splice also contributes to an increased in-fibre scattering of the source signal. The model therefore must distinguish between the theoretical, computed values and actual behavior of the fibre established by measurements.

There are other phenomenons besides the total attenuation, though. Different frequencies (or wavelengths) travel through the fibre at different speeds, therefore the light impulses received at the end of an uncompensated fibre are distorted, and care has to be taken to reconstruct the original shape of the signal before further processing [1]. There are different means of compensation for this chromatic dispersion, including Dispersion Compensating Fibre or Fiber Bragg Gratings. Properties of both must be carefully described by the model because they impact the transmission properties of the installed fibre path.

Another problem is an accurate representation of devices which are capable of changing the wavelength of the transmitted signal. Conceptually, this transformation does *not* constitute an adaptation, in the NDL/NML sense, as the signal transformation occurs on a single level of the model. However, a naive scheme which simply compared wavelength of adjacent ports of a wavelength converter would assume that a given wavelength would not be able to pass through the device.

There are other interesting phenomena, such as PMD, the polarization mode dispersion, or non-linear effects, which contribute to signal deterioration over a fibre span. An accurate description of these phenomena is needed if the model is to be used as a basis for planning and procurement of optical networks which offer a photonic service to their users.

IV. Conclusion and Future Work

This work has illustrated that while the NML model is promising, it cannot presently accommodate the requirements of a photonic service without further work. Hence, the NML scheme should be extended to allow describing the photonic layout of the network. The extended NML scheme should be self-describing, that is, it must contain enough information to allow its interpretation by a machine with no prior knowledge of photonics, but with a thorough understanding of the NML features. That way, a compliant application designed for supporting all aspects of NML will be able to work even in face of future additions to the conceptual model without a costly retrofit.

In order to verify feasibility of the chosen approach, a

 ISBN: 978-1-61208-398-8

sizable chunk of a real-world, production network should be described at the chosen level of detail. Adaptations between the low-level photonic layer and the upper layers shall be developed and tested to ensure that the NML's concepts allow an accurate representation of functionality of the modern hardware. To further prove the viability of the general concept and to study applications of semantic network modeling, a demo application shall be built. The technical demo should cover adaptations across multiple layers as well as verification of example properties of the modeled network.

The extended NML suite can be used not only for modeling of existing networks, but also for a design of new transmission paths. A user should be assisted with an extensive, semi-automatic validation of the proposed network topology. For example, while it is perfectly acceptable from the perspective of the semantic model to connect two fibres with a sequence of two add-drop multiplexers which effectively filter out all channels, the resulting topology would present little utility to its operator. A powerful validation framework should therefore be built to check whether the end product of the design is *usable*, i.e., to verify signal propagation paths across the network.

In order to prevent obsolescence and a dangerous situation where the modeled situation no longer matches reality, work should be undertaken to periodically reconcile the model with actual network in an automated manner; a topic which involves queries and call-outs to physical devices deployed in the field.

It is our plan to explore these possibilities in our work at the Optical networks department at CESNET, and in the upcoming GN4 project at GÉANT. As a first step, we will define the required NML extensions on a formal level, and verify their utility on a selected transmission path in the CESNET2 network. We also plan to use the NML extensions in formal description of the Photonic Testbed, a testbed-as-a-service (TaaS) laboratory developed at CESNET, and to benchmark its utility for describing the lowest layer of the GÉANT Testbed Service (GTS), a Europe-wide distributed testbed network.

Acknowledgements

This work was supported by the Czech institutional funding of research by project Large Infrastructure CESNET LM2010005 and by the GN3 project under the EU FP7 programme. The authors would like to thank Jan Radil and Josef Vojtěch for their valuable feedback.

References

[1] J. Kundrát, "Why we should care about the photonic layer," in 2nd TERENA Network Architects Workshop, November 2013. [Online]. Available: http://www.terena.org/activities/netarch/ws2/slides/131113-cesnet-photonics.pdf

[2] J. van der Ham, F. Dijkstra, F. Travostino, H. Andree, and C. de Laat, "Using RDF to describe networks," Future Gener. Comput. Syst., vol. 22, no. 8, Oct. 2006, pp. 862–867. [Online]. Available: http://staff.science.uva.nl/~vdham/research/publications/0510-NetworkDescriptionLanguage.pdf

[3] "Resource description framework (RDF)." [Online]. Available: http://www.w3.org/RDF/ [Accessed: 2015-03-05]

[4] J. van der Ham, P. Grosso, R. van der Pol, A. Toonk, and C. de Laat, "Using the network description language in optical networks," in Integrated Network Management, 2007. IM '07. 10th IFIP/IEEE International Symposium on, May 2007, pp. 199–205. [Online]. Available: http://staff.science.uva.nl/~vdham/research/publications/0606-UsingNDLInOpticalNetworks.pdf

[5] F. Dijkstra, "Framework for path finding in multi-layer transport networks," Ph.D. dissertation, 2009. [Online]. Available: http://www.macfreek.nl/work/Dijkstra-multilayer-pathfinding.pdf

[6] J. van der Ham, "A semantic model for complex computer networks: The network description language," Ph.D. dissertation, 2010. [Online]. Available: http://staff.science.uva.nl/~vdham/research/publications/vdham-phdthesis.pdf

[7] F. Dijkstra, "Lessons learned in multilayer network modelling," in Phosporous/Federica tutorial at TNC 2008 and NML-WG meeting at OGF 23, June 2008. [Online]. Available: http://staff.science.uva.nl/~fdijkstr/presentations/20080518%20Network%20Modelling.pdf

[8] L. Altmannová et al., "Recognizing, description, deployment and testing of new types L0/L1 resources," June 2009, pp. 17–39, deliverable Phosphorus-WP6-D.6.9. [Online]. Available: http://www.ist-phosphorus.eu/files/deliverables/Phosphorus-deliverable-D6.9.pdf

[9] "cNIS, Common Network Information Service." [Online]. Available: http://geant3.archive.geant.net/service/cnis/pages/home.aspx [Accessed: 2015-03-05]

[10] "The NML working group (NML-WG)." [Online]. Available: https://forge.gridforum.org/sf/projects/nml-wg [Accessed: 2015-03-05]

[11] "Open grid forum (OGF), an open global forum for advanced distributed computing." [Online]. Available: https://www.ogf.org/ [Accessed: 2015-03-05]

[12] G. Roberts, I. Monga, and T. Kudoh, "Network service interface," in GLIF Meeting, Atlanta, March 2014. [Online]. Available: http://www.glif.is/meetings/2014/winter/roberts-nsi-tf.pdf

[13] M. Živković, "Multi-constrained path selection," in GLIF Meeting, Atlanta, March 2014. [Online]. Available: http://redmine.ogf.org/dmsf_files/13221?download=

[14] C. de Laat and F. Dijkstra, "NML outreach session," in OGF 31, Taipei, March 2011, p. 5. [Online]. Available: http://www.delaat.net/talks/cdl-2011-03-22a.pdf

[15] J. Vojtěch, V. Smotlacha, S. Šíma, and P. Škoda, "Photonic services and their applications," in Infinity Tridentcom Thessaloniki, Greece, June 2012. [Online]. Available: http://czechlight.cesnet.cz/documents/publications/transmission-systems/2012/PS_Tridentcom12_2.pdf

[16] J. Vojtěch, V. Smotlacha, and P. Škoda, "Optical infrastructure for precise time and stable frequency transfer," Proc. SPIE, vol. 8866, 2013, pp. 886 620–886 620–5. [Online]. Available: http://czechlight.cesnet.cz/documents/publications/network-architecture/2013/ITMSfin.pdf

[17] J. Vojtěch, M. Hůla, J. Nejman, J. Radil, and P. Škoda, "Equipment for open photonic networking," in CEF Networks Workshop, 2010. [Online]. Available: http://czechlight.cesnet.cz/documents/publications/fiber-optics/2010/Vojtech-EquipmentForOpenPhotonicNetworking.ppt

 ISBN: 978-1-61208-398-8

A Peer to Peer Architecture Applied to Multiplayer Games

Felipe Rocha Wagner, Marcio Garcia Martins, Arthur Tórgo Gómez
Postgraduate Interdisciplinary Program in Applied Computing
University of Vale do Rio dos Sinos
São Leopoldo, Brazil
e-mail: feliperw@msn.com, marciog@unisinos.br, breno@unisinos.br

***Abstract*— This article presents an architecture model developed on a Peer to Peer network, which gives support to develop multiplayer games that need to manage their peers connections and permissions. The model enables the development of multiplayer games, without the need of a dedicated server, as is observed in the most architectures. For this, the model offers a library that enables programmers access the network addresses, allowing them manage their peer connections and permissions. As results, of using this architecture model, we can be cite the reduction of the costs for developers of multiplayer games due to no need a dedicated server, and a greater flexibility to manage the peer connections and permissions by the use of the available library of the model.**

Keywords-manageable network; peer to peer; network address translator; transversal problem.

I. INTRODUCTION

The multiplayer games market comes growing in the last few years. It all started with the arcades and non-networked games as Spacewar![1] and Pong [1], what later evolved to become the networked online multiplayer games that we know today. Online multiplayer makes it easy to find people to play anytime and anywhere. Nonetheless, to connect many people in order to allow them to play together, we need a server or a Peer to Peer (P2P) mesh connection. Dedicated game servers are usually expensive for indie game developers. The cost arising from the use of dedicated game servers could be reduced using a P2P approach, when designing the game network. This is one point investigated in this work. P2P networks are not easy to build; there are many technological barriers that have to be broken to connect two or more peers in different private networks. Currently, this is a challenge for multiplayer games developers. Typically, each machine in a private network is hidden behind a public gateway, a public IP, with a Network Address Translator (NAT).

In this paper, we propose a P2P architecture applied to multiplayer games that need to manage their peers connections and permissions. The idea is to create a library which allows programmers access the network addresses, without a use of a server to manage the peer connections and permissions. The admin could define users groups in accordance with the dynamic of the multiplayer games. This way, we can have an admin being responsible for the network management. This admin, in a meeting application could mute users when someone is talking or divide the meeting at a moment when needed. In the same way, this admin could manage and balance the dynamic of game rooms.

This article is structured as follows. In Section 2, related works that were utilized to generate the architecture model proposed are presented. In Section 3, we introduce the architecture model and its modules and communication protocol. Section 4 presents the peers connection process. Finally, in the Section 5, the conclusion is presented.

II. RELATED WORK

In this section, we discuss on Super Peer in P2P Networks and NAT transversal problem that were utilized to generate the architecture model proposed in this paper.

A. *Super Per in P2P Networks*

According to Yang and Garcia-Molina [2], Super Peer is a node, in a P2P network, that works both as a server to a subset of clients as a peer in a network of Super Peers. Cao et al. [3] proposed a multi-level super peer based on P2P architecture designed to work in a hierarchical structure. The hierarchical model not only distributes the single points of failure in the network, reducing the chances of presenting a massive failure, but also helps in the development of servers or applications that are based on the same model. Based on this two the related proposals, we defined a variation of Super Peer. Our Super Peer (or Admin) is a peer in the network being responsible for the network management. It can work as a server for a set of clients connects to it, and, optionally, also works as a peer to the same set of clients. The Super Peer Network, that connects Super Peers with one another, will not be considered by this model.

B. *The Network Address Translator Transversal Problem*

The NAT is a table that translates private addresses to public addresses. The development of P2P applications utilizing the NAT has constraints, because it is not possible that two or more computing systems, in different private networks, send messages between them without a public address [4]. Some techniques that allow us to break this barrier appeared along the years [5]-[12]. The most common of these is the Hole Punching, which uses discovery and prediction techniques to find out the NAT mapping. More recently, some protocols as NAT-PMP [8], PCP [9] and UPnP [11][12] utilize communication protocol to configure

 ISBN: 978-1-61208-398-8

the gateway and create a port-forwarding without the need of the user configuration.

III. ARCHITECTURE MODEL

The architecture model was designed in a way to support message packets transferred and media streams between the network's peers. It also takes into account the existence of a Super Peer, which has the ability to manage the network configuration and permissions of other peers.

Every peer has its own network module, a set of configuration flags (that describe the permissions and communication rules), an ID number and a group; the latter two are defined by the network admin.

The model proposed can be described as a hybrid model of Client-Server and P2P. The admin user initially registers himself in a server and waits for connections from common users. After the connections are made, the users communicate directly with each other, without the need of a server that would increase the costs of this process. The only function of the server is to make possible the connection of the P2P network

A. Modules

The network modules are the core of all communication. Every peer has its own module, and every module is composed by two sub-modules, as shown in Fig. 1.

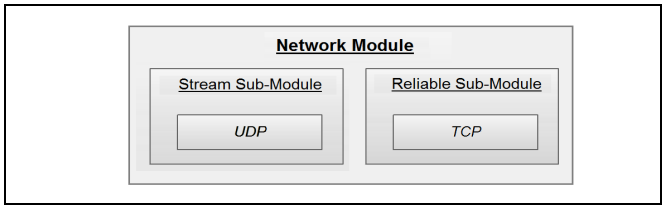


Figure 1. Network model.

The first one is a stream module for receiving and transmitting audio and video in real time using UDP, described by the RFC 768 [12]. The second one is responsible for delivering and receiving packets with network messages such as control and validation messages, and signals or important application messages. Those messages need to be sent, through a reliable connection without losing packets. Therefore, we chose to use TCP, described by the RFC 793 [13], which ensures the arrival of the packets in their destination [3][14].

To fully understand the sub-modules, we need to look at them separately. The Stream Sub-Module uses two UDP sockets, one to receive and other to transmit audio and/or video streams, as shown in Fig. 2.

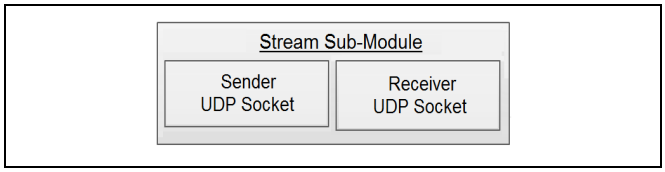


Figure 2. Stream sub-module.

On the other hand, the Reliable Sub-Module is composed of a listener responsible for receiving new connections and a list of sockets containing a functional socket for each connection sustained for peer, as seen in Fig. 3.

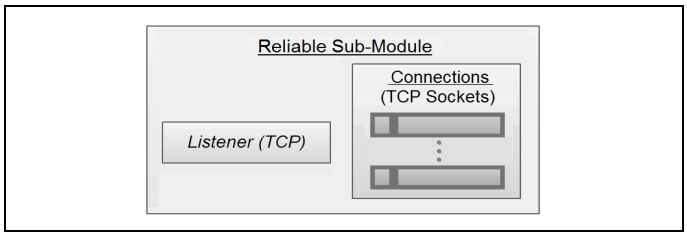


Figure 3. Reliable Sub-Module.

B. Comunication Protocol

According to Tanenbaum [15], a protocol is the set of rules and conventions governing communications between two or more computer systems. In the architecture model developed in this work, the admin can include himself in any kind of communication in his network, thus taking a full view of everything what is going on.

Users, in this architecture model, are divided into groups. Every user is connected to the admin, but not necessarily with the others users. The common users have connections with other common users, only when they belong to the same group. This way, the admin is able to send messages to any group in the network or even send a message from an user of a group to an user of another group. However, an user from a group is unable to directly send a message to an user of another group and vice versa.

In order to provide good performance to group system, the communication channels UDP and TCP have three ways to sending packets. The first one is the simple unicast, which is nothing more than the exchange of packets between two peers. The second one is multicast, and is used to send messages to a preset group of peer. At last, the third one is a broadcast, which is used to send messages to every peer with connection to the network.

The broadcast and multicast procedures can be simulated taking in account only the network mesh connections. Also, as observed in the Table I, broadcast and unicast procedures might present different behaviors according which the configurations and permissions of the peers.

TABLE I. BROADCAST & UNICAST BEHAVIOR

	Peers	Behavior
BroadcastM	Admin	Multicast → All
BroadcastG	User	Multicast* → Group
BroadcastR	User	BroadcastM Request (User → Admin)
Unicast	Admin-User	Unicast
Unicast	2 Users	Unicast or Multicast*

*Optionally might include the Admin as addressee

 ISBN: 978-1-61208-398-8

A user might broadcast messages in two ways. The first, we will call BroadcastG works as a multicast for the group the user belongs. The second occurs when a requests user to the admin to route a message for all network, including each single group, similar to a Broadcast Unknown Server (BUS) [15], that we are calling of BroadcastR.

To be able to use BroadcastR, an user must be enabled. A unicast between the common user and the admin will always be a simple unicast, but a unicast between two common users might behave as a multicast when the admin is included in the communication through the permissions and configurations of the peer. It is important to reinforce that the multicast and broadcast that we talk here might be simulated as a set of unicasts in its core.

C. *Network Packages*

The network messages can be wrapped in TCP or UDP packets, and are divided in two main groups: Common Messages and Control Messages. The difference between the two is a validation key of two bytes, appearing at the end of the packet header in the Control Messages, shown in the Table II.

TABLE II. HEADER OF WRAPPED PACKETS (EXT. = 0)

Offset (Bytes)	1 Byte			1 Byte	
	4 bits	2 bits	2 bits	2 bits	6 bits
0	Version	Type	Addressee Type	Ext. (= 0)	Reserved
2	Sender ID			Addressee ID	
4	Validation Key*				

*Present only in Control Messages

It is important to highlight that the validation key has the purpose to avoiding cheating in the network. The header starts with a four bits version number, matching the bits 0001. Next, we have two bits that define the type of message according to Table III.

TABLE III. MESSAGE TYPES

2 Bits Value	Message Type
00	Common Message
01	Common Message (Stream)
10	Control Message
11	Connection Message

The next two bits represent the addressee type, and define what will the Addressee ID corresponding, as follow. Addressee type: equal to zero (bits: 00) corresponds to a user; equal to one (bits: 01) corresponds to a group; equal to two (bits: 10) corresponds to a broadcast message; and equal to three (bits: 11) corresponds to a system message. After that, we have other two bits, which are used to establish the extension of the Sender and Addressee IDs as 2n Bytes, where n is the extension value.

The next 6 bits are reserved and should be ignored. The Bytes in sequence, should be construed according to the extension value. In case of the extension value is zero, the third Byte represents the Sender ID, and the fourth Byte represents the Addressee ID: which must be translated according to the Addressee Type value.

Only the admin has permission to send broadcast messages to the network. The users might request to the admin to send a broadcast message. If the users have the right permissions, the admin will work as a BUS sending the messages to all the users connected to him. To make a broadcast request, the common user must send an unicast message to the admin with its Addressee Type set as broadcast and the Addressee ID set to zero. It is up to the admin accepts or declines the request.

The stream transmission is equivalent to a common message, once there is no need for any validation of the frames arrival, what could cause delays in the transmission. We can stream audio, video or both (mux). To send and receive streams we must use an encoder and a decoder that will be responsible for processing the data. In this fashion, the codec or mux to be used is the responsibility of the application or of game developer.

IV. CONECTION PROCESS

To connect the peers in a network, we must follow a connection protocol. The connection protocol for this architecture model is defined in Fig. 4.

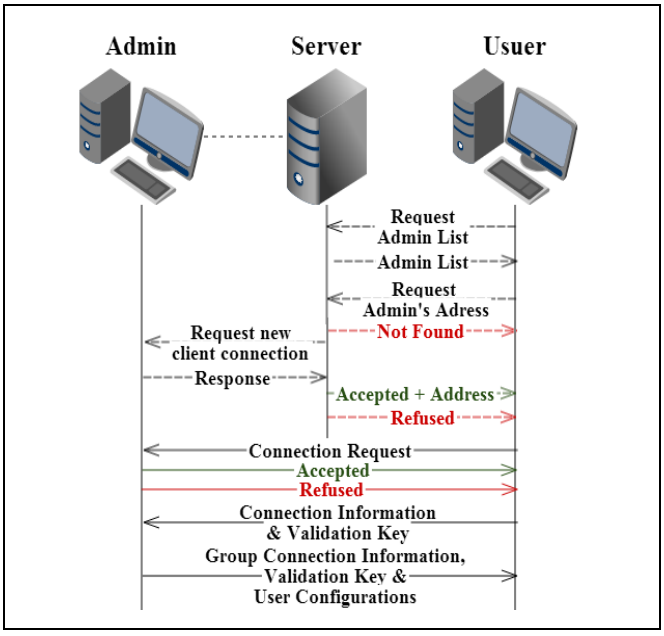


Figure 4. Connection Process.

The Admin must register its address in a server, so that common users can find him. Then, the Admin waits for connections from common users. The common user can request to server a list of registered Administrators and make a request of an address of an specific Admin. When the server receives the request, he sends a message to this

Admin requesting permission for the user to establish connection. If the user is accepted by Admin, the server sends to the user the admin address. Those messages are sent using UDP.

Once that user has the IP address and access to the Admin, he can realize the connection process. He sends a connection request over TCP to the admin, if accepted will be sent to user his UDP address, the group ID, a list with the connections information of the user from in the group, a set of flags that defines configuration and permission settings, and a two bytes validation key.

Subsequently, the address of the user is sent to the users connected to the group to which he was assigned. At the end of this process the user is added to the list of connected peers.

A. *Network Configuration Flags*

The network configuration flags describe the types of messages that will include the Admin as addressee and the permissions of each peer. Those flags' values are defined by the Admin during the connection process and are distributed over a Byte where each bit is equivalent to a Boolean that corresponds to a specific type of message. As shown in Table IV, the first four bits are related to common messages and the following four bits are related to streams. Since the control messages are always between an Admin and a common user there is no need to configure them.

TABLE IV. CONFIGURATION FLAGS

Type	1st bit	2nd bit	3rd bit	4th bit
Common	User	Group	Broadcast	System
Stream	User	Group	Broadcast	System

The bits corresponding to group messages and user, identify if those messages should include the Admin as addressee and the bits corresponding to Broadcast. The configurations of every user are saved by the Admin for validation purposes. To change the flags of a user, the Admin can send a control message with the new configurations and permissions.

V. CONCLUSION AND FUTURE WORK

In this work, we proposed an architecture model applied to a manageable P2P network that gives support to the development of multiplayer games that need to manage their peers connections and permissions. The users can communicate directly with each other, without the need of a server that would increase the costs of this process.

We presented a brief study of the NAT Transversal and some of the available techniques to break the NAT barrier in order to allow connections between hosts in different private networks. Were discussed the network model and the protocol that provides the functionalities that help both in the development of multiplayer games, as in the control of the network and in the managing of the connection processes.

As future work, we are developing a library, from the proposed architecture, in order to test the quality, usability and performance of developed applications.

REFERENCES

[1] M. Barton and B. Loguidici, "The History of Spacewar!: the best waste of time in the history of the universe," 2015 [On line]. Available from: http://www.gamasutra.com/view/feature/132438/ the history_of_spacewar_the_best.php [retrieved: Mar., 2015].

[2] B. Yang. and H. Garcia-Molina, "Designing a Super-Peer Network," Proc. International Conference on Data Engineering (wICDE), Mar. 2003, pp. 49-60, ISSN: 1063-6382.

[3] Z. Cao, K. Li, and Y. Liu, " A Multi-Level Super Peer Based P2P Architecture," Proc. International Conference on Information Networking (ICOIN), Jan. 2008, pp. 1-5, ISSN 1617-5468, ISBN 3-88579-366-0.

[4] J.F. Kurose and K.W.Ross, Computer Networking: a top-down approach, Pearson Education, 6th ed., Mar. 2012, 864 p., ISBN-13: 978-0132856201, ISBN-10: 0132856204,

[5] S. Cheshire, M. Krochmal and K. Sekar, 2006. NAT Port Mapping Protocol (NAT-PMP). [Online] Internet Draft. Available from: http://tools.ietf.org/id/draft-cheshire-nat-pmp-02.txt [retrieved: Mar., 2015].

[6] RFC 3489, 2003. STUN – Simple Transversal of User Datagram Protocol [online] RFC. Available from: http://tools.ietf.org/html/rfc3489 [retrieved: Mar., 2015].

[7] RFC 5389, 2008. Session Transversal Utilities for NAT (STUN) [Online] RFC. Available from: http://tools.ietf.org/html/rfc5389 [retrieved: Mar., 2015].

[8] RFC 6886, 2013. NAT Port Mapping Protocol (NAT-PMP) [online] RFC. Available from: http://tools.ietf.org/html/rfc6886 [retrieved: Mar., 2015].

[9] RFC 6887, 2013. Port Control Protocol (PCP) [Online] RFC. Available from: http://tools.ietf.org/html/rfc6887 [retrieved: Mar., 2015].

[10] H. Suzuki, Y. Goto, and A. Watanabe, "External Dynamic Mapping Method for NAT Transversal, " Proc. International Symposium on Communications and Information Technologies, Octo. 2007, pp. 723-728, ISBN: 978-1-4244-0977-8.

[11] UPnP Forum, 2001. Internet Gateway Device (IGD) V 1.0 [Online] UPnP Forum. Available from: http://upnp.org/specs/gw/igd1 [retrieved: Mar., 2015].

[12] UPnP Forum , 2010. Internet Gateway Device (IGD) V 2.0 [Online] UPnP Forum. Available from: http://upnp.org/specs/gw/igd2 [retrieved: Mar., 2015].

[13] RFC 768, 1980. User Datagram Protocol [Online] RFC. Available from: http://tools.ietf.org/html/rfc768 [retrieved: Marc., 2015].

[14] RFC 793, 1981. Transmission Control Protocol [Online] RFC. Available from: http://tools.ietf.org/html/rfc793 [retrieved: Mar., 2015].

[15] A. S. Tanenbaum, Computer Networks, Editora Campus, 3rd ed., 1997.

 ISBN: 978-1-61208-398-8

Performance Evaluation Methodology for Cloud Computing using Data Envelopment Analysis

Leonardo Menezes de Souza
Universidade Estadual do Ceará (UECE)
Fortaleza/CE - Brazil
Email: leonardo@insert.uece.br

Marcial Porto Fernandez
Universidade Estadual do Ceará (UECE)
Fortaleza/CE - Brazil
Email: marcial@larces.uece.br

***Abstract*—Cloud Computing is a new distributed computing model based on the Internet infrastructure. The computational power, infrastructure, applications, and even collaborative content distribution is provided to users through the Cloud as a service, anywhere, anytime. The adoption of Cloud Computing systems in recent years is remarkable, and it is gradually gaining more visibility. The resource elasticity with the cost reduction has been increasing the adoption of cloud computing among organizations. Thus, critical analysis inherent to cloud's physical characteristics must be performed to ensure consistent system deployment. Some applications demand more computer resources, other requests more storage or network resource. Therefore, it is necessary to propose an approach to performance measurement of Cloud Computing platforms considering the effective resource performance, such as processing rate, memory buffer refresh rate, disk I/O transfer rate, and the network latency. It is difficult to discover the amount of resources are important to a particular application. This work proposes a performance evaluation methodology considering the importance of each resource in a specific application. The evaluation is calculated using two benchmark suites: High-Performance Computing Challenge (HPCC) and Phoronix Test Suite (PTS). To define the weight for each resource, the Data Envelopment Analysis (DEA) methodology is used. The methodology is tested in a simple application evaluation, and the results are analyzed.**

Keywords–Cloud Computing; Performance evaluation; Methodology.

I. INTRODUCTION

The cloud computing infrastructure meets several workload requirements simultaneously, which of these are originated from Virtual Machine (VM). The evaluation addressed in this work is focused on criticality and performance on the cloud platform virtualized resources. Such evaluation is required because the performance of virtualized resources is not transparent to the network management, even when using a software monitor. Thus, it is demanded a methodology which allows to quantify the performance according to the platform particularity, using it to performance periodic measurements and to assure the promised available and reducing malfunctioning risks.

In this work, we propose a generic methodology to assess the performance of a cloud computing infrastructure; standardizing the method and covering a wide range of systems. Such methodology will serve any cloud computing structure, since it is oriented to the resources' performance. The assessment must consider the influence of each resource on the overall system performance. Then it is determined which of these resources has greater relevance to the system, aiding in deciding which infrastructure model will provide the best consumption efficiency to users, developers and managers.

We consider the average performance of the hardware and network critical points, such as processing, memory buffer refresh rate, storage Input/Output (I/O) and network latency. We used two benchmarking suites to evaluate these important points: High Performance Computing Challenge (HPCC) and Phoronix Test Suite (PTS).

HPCC uses real computing kernels, allowing variable inputs and runtimes according to system capacity [1]. It consists of seven benchmarks responsible for each critical component individual analysis according to its specificity.

The PTS [2] is the basic tool of the *Cloud Harmony* [3] website, which analyzes public cloud systems all over the world. It consists of over 130 system analysis tests, which were selected by its effective handling and compatibility of results, with higher stability and likelihood when compared to benchmarks with the same goal.

From the results obtained in both benchmark suites, we analyze it using Data Envelopment Analysis (DEA), which will assign weights according to each resource's relevance in the infrastructure; then transcribe a formulation considering each resource's average performance in each deployed VM instance. The formulation considers the overhead attached to each evaluated resource, culminating in its real performance representation. The proposal was validated in a experiment done in a Datacenter running a typical Web application.

The rest of the paper is structured as follows. In Section II, we present some related work, and Section III introduces the proposed performance evaluation methodology. Section IV shows the results and Section V concludes the paper and suggests future work.

II. RELATED WORK

Ostermann [4] and Iosup [5] created a virtual platform using the Amazon Elastic Compute Cloud (EC2) [6] instances. In this scenario, the infrastructure is shared by many independent tasks, and the benchmarks will run over the Multi-Job Multi-Instance (MJMI) sample workloads. It was noticeable two main performance characteristics: the workload makespan stability, and the resource's aquisition/liberation overhead.

The performance of several cloud computing platforms, e.g., Amazon EC2, Mosso, ElasticHost and GoGrid, were suitable to using the HPCC benchmark suite. It was noticeable that cloud computing is a viable alternative to short deadline applications, because it presents low and stable response time. It brings a much smaller delay for any cloud model when compared to scientific environment, meeting effectively to the stability, scalability, low overhead and response time criteria. The contribution of these works stands for the methodology

 ISBN: 978-1-61208-398-8

and the metrics evaluation, besides the pioneering idea of analyzing the performance of cloud computing systems [5].

Benchmark's references for performance verification and infrastructure limitations were made in [7]. The benchmarks were classified in three categories according to the moment of the infrastructure (deployment, individual or cluster). All of them brings a sense of loss carried by virtualization. In this work, it was executed simulations to assess the Central Processing Unit (CPU)/Random Access Memory (RAM), storage I/O and network usage metrics. It was verified that CPU usage tests have a little overhead introduced by virtualization. The I/O tests show performance gain caused by virtualization. Such fact possibly occurs because virtualization creates a new cache level, improving the I/O performance. On the other hand, there are components, which execute I/O functions that are affected by large cache, reducing performance and becoming the cache useless. It is difficult to predict the performance behavior in a specific I/O task.

The increasing complexity and dynamics in deployment of virtualized servers are highlighted in Huber [8]. The increasing of complexity is given by gradual introduction of virtual resources, and by the gap left by logical and physical resource allocation. The dynamics increasing is given by lack of direct control over hardware and by the complex iterations between workloads and applications. Results of experimentations using benchmarks presented that performance overhead rates to CPU virtualization is around 5%. Likewise, the performance overhead to memory (RAM), networks and storage I/O virtualizations reach 40%, 30% and 25%, respectively.

Different from cited works, this paper presents a proposal to evaluate a cloud computing system considering the application demand. Although it is possible to use HPCC or PTS metrics and calculate an index weighted by parameters based in operator experience, the results are not precise. Our proposal uses DEA methodology to define the relevance of each parameter and calculate a unique value to compare against other cloud providers.

III. A Methodology to Evaluate the Performance of a Cloud Computing System

Amazon Elastic Compute Cloud (Amazon EC2) is a service provided by Amazon cloud computing platform. The users can access the platform by the Amazon Web Services (AWS) interface. Amazon's offer the Amazon Machine Image in order to create a Virtual Machine (VM), which is called an *instance*, containing user's software. A user can create, deploy, and stop server instances as needed. They pay the service by the amount of hours of active server instance it used.

In each Amazon's VM, or VM instance, works as a virtual private server. To facilitate for user to choose the amount of resources they would buy, Amazon defines a set of instance size based on Elastic Compute Units. Each instance type offers different quantity of memory, CPU cores, storage and network bandwidth. The Amazon's pre-defined VM types used in this work are shown in Table I.

First, we deploy VMs based on the model provided by Amazon EC2 [6]. The overall performance of the resources is not used, since virtualization generates communication overhead in the resource management. After the allocation of resources in need, we installed the benchmark suites to run the tests.

TABLE I. Amazon EC2 Virtual Machine Model [6].

MVs	ECUs(Cores)	RAM[GB])	Arq[bit]	Disk[GB]
m1.small	1 (1)	1,7	32	160
c1.medium	5 (2)	1,7	32	350
m1.large	4 (2)	15	64	850
m1.xlarge	8 (4)	15	64	1690
c1.xlarge	20 (8)	7	64	1690

According to Jain [9], the confidence interval only applies to large samples, which must be considered from 30 (thirty) iterations. Therefore, we ran the experiments for each resource of each VM instance at least thirty times, ensuring the achievement of a satisfactory confidence interval (95%). Then, we can state that each benchmark will follow this mandatory recommendation to achieve an effective confidence interval. After the tests, we calculate the mean and the confidence interval of the obtained results, presenting a high reliability level.

In order to ponder the performed experiments, we opted for the DEA methodology; using the BCC model output-oriented (BCC-O), which involves an alternative principle to extract information from a population of results. Then, we determine the weights inherent to the VMs and the resources analyzed. We used the results of each benchmark iteration in each VM as an input, achieving the weights for each benchmark. Finally, we apply this procedure in the formulation which will be detailed later.

In short, we analyze a cloud performance simulating the behavior of applications by running benchmarks. We did an efficiency analysis from the achieved results, assigning weights to each one of them. Then, we proposed a formulation which showed the consumption ratio of each platform resource, considering the associated overhead. The execution order of activities is shown in Figure 1.

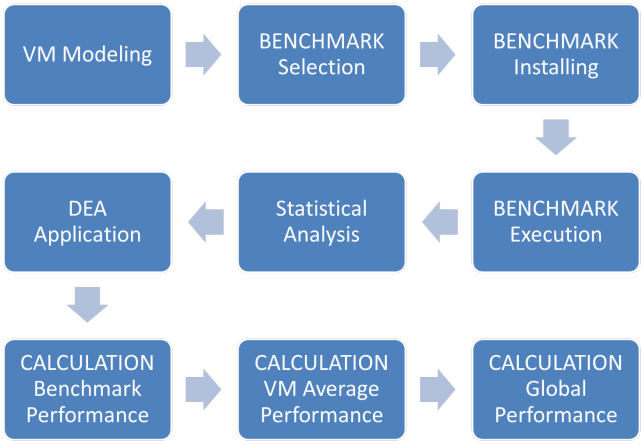


Figure 1. Cloud Computing performance evaluation methodology flowchart.

A. Benchmarks

In this work, we use two benchmark suites, the HPCC [1] and PTS [2]), which will measure the performance of critical points in a cloud computing system. These benchmarks require the Message Passing Interface (MPI) [10] and Basic Linear Algebra Subprogram (BLAS) [11] library's availability to run

 ISBN: 978-1-61208-398-8

the tests. The benchmarks from HPCC suite ran both in local and online environments and has shown favorable results to its utilization. Then, the benchmark results showed independence and adaptability within the cloud nodes.

The HPCC benchmark suite comprises seven different tests that will stress the system hardware critical points such as is presented as follows:

- High-Performance Linpack (HPL) [12] uses 64-bit double precision arithmetics in distributed memory computers to measure the floating point rate of execution for solving matrices through random dense linear equations systems.
- Double-precision General Matrix Multiply (DGEMM) [13] simulates multiple floating point executions, stressing the process through double-preciosion matrix multiplication.
- PTRANS [14] has several kernels where pairs of processors communicate with each other simultaneously, testing the network total communication capability. It transposes parallel matrices and multiplies dense ones, apllying interleaving techniques.
- Fast Fourier Transform (FFT) [15] measures the floating point rate through unidimensional double-precision discrete Fourier transforms (DFT) in arrays of complex numbers.
- STREAM [16] measures the memory bandwidth that supports the processor communication (in GB/s). It also measures the performance of four long-vector operations. The array is defined to be larger than the cache of the machine which is running the tests, privileging the memory buffer updates through inter-dependence between memory and processor.
- Random Access [17] measures the performance of random memory (main) and access memory (cache) buffer updates in multiprocessor systems. The results are given in Giga Updates Per Second (GUPS), calculated by updated memory location identification in one second. This update consists in a Read-Modification-Write (RMW) operation controlled by memory buffer and the processor.
- Effective Bandwidth Benchmark (b_{eff}) [18] measures the bandwidth efficiency (effective) through estimated latency time for processing, transmission and reception of a standard message. The message size will depend on the quotient between memory-processor ratio and 128.

Beyond the HPCC, we also used another benchmark suite to run the remaining tests and enable a bigger coverage of evaluated resources. The PTS suite comprises more than 130 system analysis tests. We have selected the benchmarks to be part of this experiment according to its importance within the benchmarking set, minimizing inconsistencies and improving our sample space. Finally, we achieve the three most adaptive benchmarks which will be presented as follows:

- Loopback Transmission Control Protocol (TCP) Network Performance [19] is a simple Peer-to-Peer (P2P) connectivity simulation which measures the network adapter performance in a loopback test through the TCP performance. This test is improved on this benchmark to transmit 10GB via loopback.
- RAM Speed SMP [20] measures the performance of the interaction between cache and main memories in a multiprocessor system. It allocates some memory space and starts a write-read process using 1Kb data blocks until the array limit, checking the memory subsystem speed.
- PostMark [21] creates a large pool of little files constantly updating just to measure de workload transaction rate, simulating a big Internet e-mail server. The creation, deletion, read, and attaching transactions have minimum and maximum sizes between 5Kb and 512Kb. PostMark executes 25.000 transactions with 500 files simultaneously, and after the transactions, the files are deleted, producing statistics relating its contiguous deletion.

In short, we present all benchmarks used in this work and its basic characteristics on the Table II.

TABLE II. BENCHMARKS CHARACTERISTICS.

RESOURCE	BENCHMARK	UNIT
CPU	HPL DGEMM PTRANS	GFLOPs
	FFT	GB/s
MEM	STREAM RAM Speed SMP	GB/s
	Random Access	GUPS
STO	PostMark	Transactions/s
NET	b_{eff}	μs
	Loopback TCP	s

B. Resources Overhead

Simplifying the organization of the resources' performance analysis in a cloud computing system, we can split them into two requirement groups: CPU and I/O resources. Performance studies utilizing general benchmarks show that the overhead due to CPU virtualization reach 5% as was mentioned before at Section II. The host hypervisor directly controlling the hardware and managing the actual operational system, showing low overhead.

Virtualization also imposes I/O overhead, concerning memory, networks and storage. Cloud applications have specific requirements, according to their main goal. In this way, the network is critical to every single cloud application because it determines the speed with which each remaining I/O resource will work. In other words, the network must provide capability, availability, and efficiency enough to allocate resources without compromising delays.

The online content storage is just one of the most popular features of cloud computing systems. Its performance is so much dependent on memory buffer updates rate as regarding the processing rate that feeds the buffer. These two active functions affect significantly the storage services on the cloud.

Lastly, but not less important, memory is the most required resource on a cloud computing system. In distributed systems, it is considered a critical issue, because it works along with processing in the updates of running applications, user requirements, and in the data read/write coming through network adapter or storage component. So, many functions overload the resource, representing the biggest bottleneck in whole cloud computing infrastructure.

 ISBN: 978-1-61208-398-8

TABLE III. VIRTUALIZATION OVERHEADS [8].

RESOURCE		OVERHEAD (%)
I/O	Memory	40
	Network	30
	Storage	25
CPU	Processing	5

Each hardware resource available in the cloud computing infrastructure possesses a unique utilization quota regarding its own functioning. However, they feature interdependencies between to each other. Table III shows the overhead portions to each resource analyzed in this work. Then, we address weights based on the significance of each resource in a cloud computing infrastructure using the DEA methodology.

C. DEA Methodology

The DEA methodology is a linear programming mathematical technique which consists of a multicriteria decision support, analyzing multiple inputs and outputs simultaneously. In this way, the DEA is capable of modeling real-world problems meeting the efficiency analysis [22].

This methodology provides comparative efficiency analysis from complex organizations obtained by its unit performance revelation so that its reference is obtained by the observation of best practices. The organizations once under DEA analyses are called Decision Making Unit (DMU)s and must utilize common resources to produce the same results. With this, will be defined efficient DMUs (those which produce maximum outputs by inputs) and the inefficient ones. The first ones are located on the efficiency frontier while the later ones under that same frontier.

In this work, we chose one model among all DEA methodology models, which is the Multipliers BCC-O model. The output orientation was chosen because of the input variables (VM instances) are fixed. The main goal is to obtain the best benchmarks' performance executed on the VMs, then we intend to obtain the larger amount of outputs by inputs. By the way, the DEA methodology was applied to parametrize the benchmarks results calculated for each resource in all VM instances.

The required terms to the weighting on the proposed formulation are generated by the BCC-O model. This mathematical model consists of the calculation of the input (VM resources) and output (benchmarks results) variables weights. In the model objective function we minimize the input weighted sum (product from input value by its respective weight) subjected to four restrictions, presented on the formulation shown in (1).

Running the model shown earlier in a linear programming solver, we can get the weight sum equal to 1, showed in (1b). The restriction of the inequality (1c) will be performed for each one of the 1500 total iterations from running instances. This model allows weights to be chosen for each DMU (VM iteractions) in a way that suits it better. The calculated weights must be greater than or equal to zero as it is shown on inequalities (1e) and (1f). The efficiency ratios of each DMU is calculated by the objective function too. Thus, the number of models to be solved is equal to the number of problem DMU.

In order to achieve the best performance of the resulting benchmarks (outputs) ran on the five VMs showed in Table I. The weights are obtained by a weighted average according to the significance of each test on the system. The greater values will have the higher weights. We consider each one of the ten benchmarks executed ran, at least, 30 times for each one of the five VMs used in this experiment, accounting for 1500 iterations. Each one of these had its respective weight calculated by DEA, then we ran a solver (BCC-O) to calculate the inputs and outputs weighted sum obeying the methodology constraints.

$$\text{Minimize} \quad ef(0) = \sum_{i=1}^{m} v_i X_{i0} + v \tag{1a}$$

$$\text{Subject to} \sum_{j=1}^{S} u_j Y_{j0} = 1 \tag{1b}$$

$$\sum_{j=1}^{S} u_j Y_{jk} - \sum_{i=1}^{m} v_i X_{ik} - v \leq 0 \tag{1c}$$

$$k = 1 \ldots n \tag{1d}$$

$$u_j \geq 0, \forall j \tag{1e}$$

$$v_i \geq 0, \forall i \tag{1f}$$

Where: $v \in \Re$, v unrestricted
u_j = output j weight
v_i = input i weight
$k \in \{1 \ldots n\}$ DMUs
$j \in \{1 \ldots s\}$ outputs of DMUs
$i \in \{1 \ldots m\}$ inputs of DMUs
Y_{jk} = output j value of DMU k
X_{ik} = input i value of DMU k

Concerning the constraints, first of all, the outputs' weighted sum must be equal to one, setting a parameter for assigning weights in each VM. The inputs and outputs' weights must be greater than or equal to zero. Lastly, the subtraction between the inputs and outputs' weighted sums and the scale factor, must be lower than or equal to zero. The scale factor will not be considered because it will just determine if the production feedback is increasing, decreasing or constant to a set of inputs and products. This way, weights are the factors considered on the formulation.

D. Formulation

In a cloud computing system, the required resources are allocated automatically according to user needs. All of them have a standard overhead and significance variable level according to hosted application guidance. To analyze the system performance, we used a mathematical formulation that provides evidence from utilization levels measured, and from the iteractions among resources. The DEA was used to define the weights of Performance Index.

We must consider that benchmark execution will simulate an application that overloads the assessed resource. Then, we adopted PI_{R_G} as the Resource Global Performance Index, whose variable will assume the resulting value from the product between RPI_R (Resource Real Performance Index) and the API_{R_j} (Average Performance Index by Resource in each VM Instance), as shown in (2).

$$PI_{R_G} = RPI_R \times API_{R_j} \tag{2}$$

 ISBN: 978-1-61208-398-8

The term RPI_R is the result from the subtraction between the maximum theoretical performance (100%) and the overhead associated to each running resource, shown on the Table III. The relation is shown in (3).

$$RPI_R = (100\% - Ov_R\%) \tag{3}$$

The term API_{R_J} is calculated by the mean of each BPI_{R_j} (Benchmark Performance Index by Resource in each Instance), as it is shown in (4). BPI_{R_j} is calculated by the product sum between weights (U_{iR_j}) obtained from DEA methodology for benchmarks (i) by resource (R) in each instance (j). The term n_j stands for the amount of VMs where benchmarks were hosted. In this case, five VMs were implemented to run the tests based on the Amazon EC2 infrastructure.

$$API_{R_j} = BPI_{R_j} \div n_j \tag{4}$$

The results (X_{iR_j}) obtained from benchmarks (i), by resource (R) in each instance (j), as shown in (5), where p is the number of benchmarks and q is the number of instances. The X_{iR_j} is normalized related to maximum theoretical performance in order to permit an index independent from benchmark units, e.g., GB/s, GFLOPS, Sec.

$$BPI_{R_j} = \sum_{\substack{1 \le i \le p \\ 1 \le j \le q}} (U_{iR_j} \times X_{iR_j}) \tag{5}$$

The benchmark suites were set up to simulate each resource behavior in a cloud computing infrastructure. We will calculate the (BPI_{R_j}) Benchmarks Performance Index to each resource (R) in each instance (j), considering each benchmark running to its respective resource, and after that we calculated the mean for each resource, obtaining the API_{R_j} dividing each BPI_{R_j} by the number of VM instances n_j. In following formulation, CPU means computing resource, MEM means memory, STO means storage resource and NET means network resource.

$$\begin{aligned}
BPI_{CPU_j} &= (U_{HPL} \times X_{HPL}) + (U_{DGEMM} \times X_{DGEMM}) \\
&\quad + (U_{FFT} \times X_{FFT}) + (U_{PTRANS} \times X_{PTRANS}) \\
BPI_{MEM_j} &= (U_{STREAM} \times X_{STREAM}) + (U_{RA} \times X_{RA}) \\
&\quad + (U_{RSMP} \times X_{RSMP}) \\
BPI_{STO_j} &= (U_{BB} \times X_{BB}) + (U_{PM} \times X_{PM}) \\
BPI_{NET_j} &= (U_{BE} \times X_{BE}) + (U_{LTCP} \times X_{LTCP})
\end{aligned}$$

$$\begin{aligned}
API_{CPU_j} &= \sum BPI_{CPU_j} \div n_j \\
API_{MEM_j} &= \sum BPI_{MEM_j} \div n_j \\
API_{STO_j} &= \sum BPI_{STO_j} \div n_j \\
API_{NET_j} &= \sum BPI_{NET_j} \div n_j
\end{aligned}$$

The next step consists in solving the global performance expression:

$$\begin{aligned}
PI_{CPU_G} &= RPI_{CPU} \times API_{CPU_j} \\
PI_{MEM_G} &= RPI_{MEM} \times API_{MEM_j} \\
PI_{STO_G} &= RPI_{STO} \times API_{STO_j} \\
PI_{NET_G} &= RPI_{NET} \times API_{NET_j}
\end{aligned}$$

IV. Results and Discussion

All the results are based on the initial set of benchmarks showed in Section III. As we could see in Table I, we created a homogeneous environment from 1 to 21 cores based on five Amazon EC2 instances, where we run the benchmarks which will evaluate the performance on the cloud environment.

The hardware used was a Dell Power Edge M1000e enclosure with six blades powered by Intel Xeon x5660 2.8 GHz processor and 128 GB 1333 MHz DDR3 RAM. All blades have 146 GB SAS HDs. The storage was a Dell Compellent with six 600 GB SAS disk and six 2.0 TB NL-SAS disk. The OS was the Linux Ubuntu 12.04 over VMWare ESXi 5.0.0 hypervisor.

The application chose was an XAMPP 1.8.1 Web server [23]. After running each benchmark, we generate Table IV which shows the efficiency index of each experiment related to maximum theoretical performance. The normalization is necessary to compare different units from benchmarks. Then, we calculated its efficiency percentage to use it on the proposed formulation.

In order to consider the results from the benchmark experiments, we used DEA methodology through BCC-O model (output-oriented). Beyond the efficiency index calculation, we calculate the output variable weights (benchmark results). In this way, we minimize the inputs weighted sum dividing it by the outputs' weighted sum of the benchmark at hand. After that, we ran a BCC-O solver to address weights to each benchmark, considering each VM instance according to its influence in the obtained results shown in Table IV. Table V shows the weights calculated by the BCC-O solver that will influence the performance of each resource attached to each benchmark in each VM.

The benchmark results were shown in Table IV and the efficiency index were calculated by DEA methodology (BCC-O) in Table V. Applying these results on (5), its two factors will assume values for benchmark performance to each resource in each instance (X_{iRj}), considering the DEA assigned weight to each benchmark result (U_{iRj}). We can observe the more the resource is used, greater is the weight assigned to it.

We can see in Figure 2 the network performance is clearly greater than the rest, and the memory is the only resource that has an index relatively close. These resources are the most affected ones by the overhead issue, justifying their bottleneck condition. The Figure 3 shows the relevance of each instance through benchmark execution. The c1 instances have very similar performances because they both have a processor/memory ratio which allows achieving quite satisfying performance levels.

From these results we verified, the memory and network performances are the most relevant to a cloud computing system. These two resources, when well balanced, leverage the cloud computing infrastructure managing workloads, reaffirming its bottleneck condition. In this way, this proposal gives more information regarding resource performance relevance in application when comparing to the work of Huber [8].

 ISBN: 978-1-61208-398-8

TABLE IV. BENCHMARK RESULT FOR EACH VM (X_{iRj}) RELATED TO MAXIMUM THEORETICAL PERFORMANCE.

BENCHMARKS		m1.small	c1.medium	m1.large	m1.xlarge	c1.xlarge
CPU	HPL	4.64%	11.27%	14.84%	24.81%	27.51%
	DGEMM	1.15%	13.27%	4.30%	8.54%	11.08%
	FFT	0.94%	3.62%	2.49%	4.52%	4.59%
	PTRANS	6.83%	27.86%	14.71%	39.63%	38.52%
MEM	RAMSpeed SMP/Integer	22.01%	28.38%	25.7%	30.4%	30.77%
	RAMSpeed SMP/Float	24.46%	28.96%	26.30%	27.13%	31.36
	STREAM	19.53%	28.27%	44.02%	37.53%	41.36%
	RandomAccess	0.41%	9.82%	3.73%	17.3%	17.6%
NET	b_{eff}	98.2%	99.9%	98.8%	99.5%	99.4%
	Loopback TCP	0.58%	62.07%	92.65%	94.34%	96.02%
STO	PostMark	3.75%	4.42%	13.99%	13.00%	14.26%

TABLE V. WEIGHTS ADDRESSED TO RESOURCES TO EACH VM (U_{iRj}).

BENCHMARKS		m1.small	c1.medium	m1.large	m1.xlarge	c1.xlarge
CPU	HPL	0.77	0.13	0.66	0.28	0.51
	DGEMM	0.88	0.42	0.23	0.29	0.20
	FFT	0.003	0.58	0.25	0.5	0.58
	PTRANS	0.15	0.32	0.07	0.33	0.43
MEM	RAMSpeed SMP/Integer	0.38	0.78	0.17	0.42	0.37
	RAMSpeed SMP/Float	0.46	0.96	0.3	0.68	0.65
	STREAM	0.14	0.18	0.67	0.33	0.61
	RandomAccess	0.91	0.93	0.37	0.73	0.56
NET	b_{eff}	0.42	0.59	0.48	0.55	0.43
	Loopback TCP	0.24	0.43	0.33	0.28	0.48
STO	PostMark	0.57	0.24	0.19	0.42	0.62

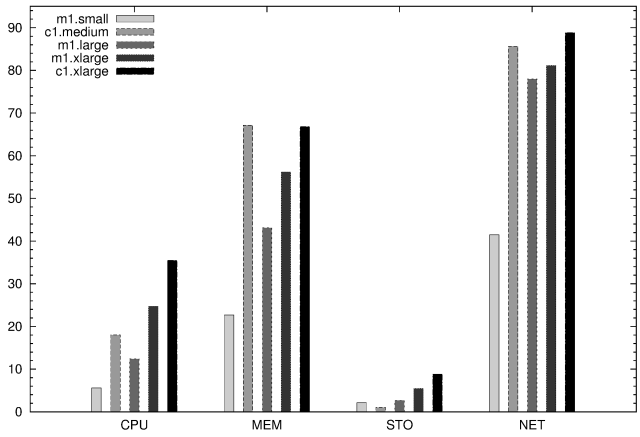


Figure 2. Benchmark Performance by Resource.

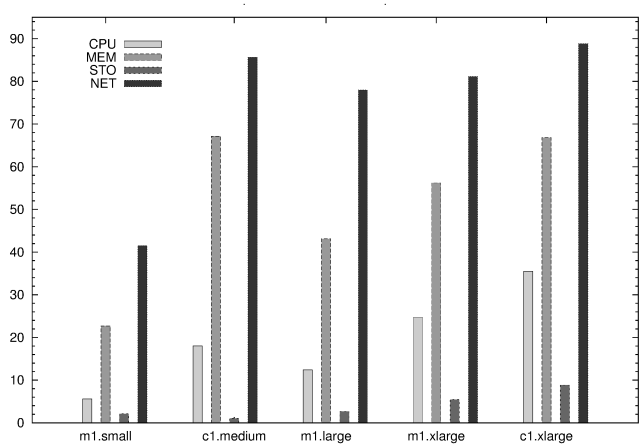


Figure 3. Benchmark Performance by Instance.

V. CONCLUSION AND FUTURE WORK

In this work, we could observe that the benchmarks had met the simulation needs very well, overloading the resources efficiently, returning real-world results. The DEA methodology helped us to analyze the efficiency of each experiment, providing an efficiency index (weight) to benchmarks in each instance implemented, for each resource evaluated. Finally, the proposed formulation highlighted the impact of resource's overhead on the global performance evaluation.

Then, we concluded that, in a generic Web application, the memory and network resource performance is the most relevant to a cloud computing system, and for this reason, they are considered the bottlenecks. We confirmed that the resource performance evaluated here is directly proportional to the overhead execution rates, assigned in [8].

Since develop an application to be hosted on a cloud environment to measure its resource consumption rate, or its behavior during a VM migration process, until configure the benchmarks in a more aggressive way, generating more data blocks. We should, then, pay attention to cloud computing system constant evolution to make possible the use of the approach proposed in this work.

REFERENCES

[1] J. Dongarra and P. Luszczek, "HPCC High Performance Computing Challenge," Last accessed, Mar 2015. [Online]. Available: http://icl.eecs.utk.edu/hpcc

[2] M. Larabel and M. Tippett, "Phoronix Test Suite," Last accessed, Mar 2015. [Online]. Available: http://www.phoronix-test-suite.com

[3] J. Read, "Cloud Harmony: Benchmarking the Cloud," Last accessed, Mar 2015. [Online]. Available: http://www.cloudharmony.com

[4] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "An early performance analysis of EC2 cloud computing

 ISBN: 978-1-61208-398-8

services for scientific computing," Cloud Computing, 2010, pp. 115–131.

[5] A. Iosup, S. Ostermann, M. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 6, 2011, pp. 931–945.

[6] Amazon, "Amazon Elastic Compute Cloud EC2," Last accessed, Mar 2015. [Online]. Available: http://aws.amazon.com/ec2

[7] N. Cardoso, "Virtual clusters sustained by cloud computing infrastructures," Master's thesis, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, Dec 2011.

[8] N. Huber, M. von Quast, M. Hauck, and S. Kounev, "Evaluating and Modeling Virtualization Performance Overhead for Cloud Environments," in 1st International Conference on Cloud Computing and Services Science, 2011, pp. 7–9.

[9] R. Jain, The Art of Computer Systems Performance Analysis. John Wiley & Sons, 2008.

[10] J. Dongarra, R. Hempel, T. Hey, and D. Walker, "The Message Passing Interface (MPI) Standard," Last accessed, Mar 2015. [Online]. Available: https://mcs.anl.gov/research/projects/mpi

[11] C. Lawnson, R. Hanson, D. Kincaid, and F. Krogh, "Basic Linear Algebra Subprograms for FORTRAN Usage," ACM Transactions on Mathematical Software (TOMS), vol. 5, no. 3, 1979, pp. 308–323.

[12] A. Petitet, R. Whaley, J. Dongarra, and A. Cleary, "HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers," Last accessed, Mar 2015. [Online]. Available: http://netlib.org/benchmark/hpl/

[13] J. Dongarra, I. Duff, J. Croz, and S. Hammarling, "Subroutine DGEMM," Last accessed, Mar 2015. [Online]. Available: http://www.netlib.org/blas/dgemm.f

[14] T. Hey, J. Dongarra, and H. R., "Parkbench Matrix Kernel Benchmarks," Last accessed, Mar 2015. [Online]. Available: http://www.netlib.org/parkbench/html/matrix-kernels.html

[15] M. Frigo and S. Johnson, "benchFFT," Last accessed, Mar 2015. [Online]. Available: http://www.fftw.org/benchfft/

[16] J. McCalpin, "STREAM: Sustainable Memory Bandwidth in High Performance Computers," Last accessed, Mar 2015. [Online]. Available: http://www.cs.virginia.edu/stream/

[17] D. Koester and B. Lucas, "Random Access," Last accessed, Mar 2015. [Online]. Available: http://icl.cs.utk.edu/projectsfiles/hpcc/RandomAccess/

[18] R. Rabenseifner and G. Schulz, "Effective Bandwidth Benchmark," Last accessed, Mar 2015. [Online]. Available: https://fs.hlrs.de/projects/par/mpi//b_eff/

[19] M. Larabel and M. Tippett, "Loopback TCP Network Performance," Last accessed, Mar 2015. [Online]. Available: http://openbenchmarking.org/test/pts/network-loopback

[20] R. Hollander and P. Bolotoff, "RAMspeed," Last accessed, Mar 2015. [Online]. Available: http://alasir.com/software/ramspeed/

[21] J. Katcher, "PostMark: A New File System Benchmark," Last accessed, Mar 2015. [Online]. Available: http://www.netapp.com/technology/level3/3022.html

[22] W. W. Cooper, L. M. Seiford, and K. Tone, "Data envelopment analysis: A comprehensive text with models, applications, references and dea-solver software. second editions," Springer, ISBN, vol. 387452818, 2007, p. 490.

[23] K. Seidler and K. Vogelgesang, "XAMPP Distribution Apache + MySQL + PHP + Perl," Last accessed, Mar 2015. [Online]. Available: https://www.apachefriends.org

 ISBN: 978-1-61208-398-8

NDNGame: A NDN-based Architecture for Online Games

Diego G Barros

Universidade Estadual do Ceará (UECE)
Fortaleza-CE, Brazil
Email: diego.barros@larces.uece.br

Marcial P Fernandez

Universidade Estadual do Ceará (UECE)
Fortaleza-CE, Brazil
Email: marcial@larces.uece.br

***Abstract*—The content-oriented network paradigm is an alternative approach for computer's networks, proposing a new communication architecture compatible with the dynamic nature of the current Internet. Among these models, we can mention Named Data Network (NDN). Its basic idea is to retrieve data through content names, instead of source and destination IP address. Using in-network caches, this approach allows to achieve good performance to distribute content in large-scale, improving the usage of the network. However, this model is not a consensus on end-to-end applications such as e-mail, VoIP, games and client-server application. The NDN protocol overhead reduces the performance for these applications. This work proposes a hybrid network architecture in online games, using NDN for content dissemination and point-to-point IP communication to deliver control messages. Our proposal demonstrates how NDN networks can be used to improve online game's distribution network maintaining the user experience.**

Keywords–Future Internet; Online game architecture; Named Data Network (NDN).

I. INTRODUCTION

The Internet project was made 50 years ago, basically, focusing on point-to-point communication and technical users. Due to the telecommunication evolution and popularization of computers, the Internet became a successful, effective, global-scale communication.

As a consequence, new service demand and products were offered, like, e-commerce, social networks, file sharing, Voice over Internet Protocol (VoIP), online games, video stream, and others. However, to make this possible, the Internet's architecture suffered many amendments, the Internet infrastructure becoming more complex, increasing the cost for implementation, maintenance and management of these applications. This process is known as the Internet ossification [1].

Games are popular applications, and deliver a huge amount of multimedia content. The data flow in merely one game distributor [2] can reach 13.2 PB per week. Only in USA [3], 59% of North-Americans play some video game, spending in 2013 a total US$ 21.53 billions, out of which US$ 15.39 billions was just for content purchase. According to gamer's company demand, we believe Named Data Network (NDN) architecture could support games provider needs.

Online games impose a significant challenge in current Internet architecture. The huge amount of data delivered as scenario, video and images, and the necessity to provide a fast response to game commands brings new challenges to network researchers. It is easy to guarantee reduced packet delay for low bandwidth application. However, when we consider high bandwidth, the Quality of Service (QoS) guarantee becomes more difficult.

The online game application produces an incompatibility between models, the original Internet's architecture and current applications. IP packets were predicted for simple end-to-end applications, but the dynamic nature of the Internet requires more flexibility. The great content production, 500 exabytes in 2008 [4], is only one example. Considering the Internet growth rate, we estimate more than 1.5 zettabyte in 2014.

Today, the websites are evaluated by "what" content they contains. However, the Internet communications works in terms of "where" the content is. Then, the current architecture causes incompatibility issues to new applications, as availability, security and local-dependence.

Due to this scenario, some researches propose to reorganize the Internet's architecture. These proposals are divided into two types. The first type proposes small incremental changes, while the second proposes a redesign from scratch, changing the core principles. The second approach is known as Clean Slate [5].

Among the different approaches, we can highlight the content-based networks. In 2009, Palo Alto Research Center (PARC) defended a proposal Content-Centric Network (CCN) that today is known as NDN. The basic idea is to retrieve content by the name, instead of origin and destination address [6]. This approach has a new communications principle that improves abstraction and performance of networks.

Basically, a NDN node attends content requests through data sharing. This model supposes that each node can provide caching service, according to its resource dependence and policy.

Within this proposal, the NDN shows simplicity and flexibility with similar functions of current networks. However, among many advantages to this model, its main virtue is to improve content distribution. The NDN works on demand, improving performance and scalability. Thus, services with great content dissemination, e.g., video stream and online games, will be benefited.

However, NDN model is not well suitable for point-to-point applications. It is not clear how it provides efficiently traditional applications such as VoIP, e-commerce, online games, e-mail. Then, a pure deployment of NDN networks it is very improbable. Thus, we believe that the best way to deploy NDN is through an overlay network, e.g., torrent application to share content over IP infrastructure. Therefore, it is significantly important to validate the NDN approach to provide a good performance in generic web applications.

To overcome this challenge, this paper proposes a Hybrid Network Architecture for online games, using NDN for content dissemination and point-to-point IP communication to deliver

 ISBN: 978-1-61208-398-8

control messages. Our proposal demonstrates how NDN networks can be used to improve online game's distribution network maintaining the user experience. A prototype over Mininet tool [7] running Quake 3 game server to evaluate the proposal was developed.

The rest of the paper is structured as follows. In Section II, we present some related work about real time application on content networks. Section III introduces the NDN concepts and Section IV presents the online game architecture. In Section V, we present the NDNGame Architecture, and its main blocks. Section VI shows the proposal evaluation, and finally, in Section VII, we conclude the paper and present some future work.

II. Related Works

A. Donny Brook

Donny Brook [8] is a game architecture based on a peer-to-peer model to run Quake 3. It aims to improve bandwidth, reducing the set of interest objects, as consequence, promoting a significant decrease in updating messages. Another relevant contribution is the use of a multicast message system, allowing multiple updating sources, sensitive to the response time. It also implements a load balance mechanism, where powerful nodes can support others. This work is important to our proposal, because it evaluates a valid Quake 3 implementation with similar goals, which is used as an alternative client-server architecture with enhanced performance using full advantage of network.

B. Voice over CCN

There is a meaningful importance in validating point-to-point applications, like email, VoIP, in Content-Oriented Network [9]. Jacobson et al. adapt a VoIP application to a content network [10]. They use an "on demand" publishing system, which serves as a contact point to the service, allowing users to initiate the session. Due to the use of names on NDN networks instead of IP's addresses, it was introduced the concept of constructable names, where is possible to build names of desirable contents without having seen the exact content name before. Thus, with a deterministic algorithm, the consumer and the producer can retrieve the same name on information available to both. This work serves as main reference to the use of a point-to-point application in NDN.

C. G-COPSS

GCOPSS is a distributed game platform that uses a Content-Oriented Network to deliver objects [11]. It adapts the COPSS [12] to improve scalability, which is an important goal on game's environments. It discovers network topology in order to offer an efficient system to disseminate the content. G-COPSS uses a hierarchical content descriptor and also implements a framework to provide content dissemination based on publishing requests. A user expresses interest in Content Descriptors (CD), e.g., /sports/soccer. The content publishers send announcements related to a specific CD when new parts of the content arrive. CDs are hierarchically organized. High-level users can receive announcements from users in a different level (lower), e.g., /sports receives /sports/soccer or /sports/swimming. NDN requires a new forward engine. The routers implement a Subscription Table (ST). STs maintain a CD base with subscriber's information, working in a distributed manner, as well using IP multicast to deliver content. This work shows the updating message exchange on an online game network using the NDN paradigm.

D. MERTS

More Efficient Real-time Traffic Support (MERTS) reinforces the importance to optimize NDN networks for point-to-point applications. Video and audio stream are much more sensible to network delay, thus, MERTS proposes a content classification in real-time and not real-time for on demand traffic. However, it is necessary to add a new field in NDN packets, modifying its basic structure. Our approach does not impose any modification on NDN design; instead, we propose a modification at the application layer, maintaining the NDN architecture.

III. Named Data Networks

The basic idea about Content-Oriented Network is not new; research like TRIAD project in 1999 and Data-Oriented Network Architecture (DONA) in 2006, already used content object name to forward packets [13]. In 2009, the PARC group published the proposal of content-centric architecture, which then became known as NDN [9]. Nevertheless, the NDN model stands out since it does not need an origin and destination address like IP on traditional networks. Therefore, an NDN network requires only the content name to retrieve it. This philosophy is simple and it can solve many problems like availability, security and location-dependence.

In order to understand this subject, it is necessary to understand how an NDN network works. However, before defining an NDN node, it is important to know that it works basically with only two packet types: Interest packet and Data packet. When a consumer needs a content, he expresses it by content name in the Interest packet; this packet is sent via broadcast over all network connections. The Data packet is the content which attends the Interest packet. Technically, it only occurs when both possess the same Content Name NDN works with the "face" concept, which is a reference of the requested origin, and it may be anything as an IP address, MAC, proxy, application, among others.

A NDN node is composed by the following entities:

1) Content Store: it works basically like a content buffer memory, storing the content disseminated by the network, but with a distinct replacement policy. NDN packets have an idem potent property, for different requests it may return the same result, like a Youtube video can satisfy user A, as well user B, C and D.
2) Pending Interest Table (PIT): it is essentially the table of interests not attended yet. When an interest is disseminated on NDN network, the correspondent PIT table in each node stores the interest name, and the face which it was requested. So, when a content matches the interest, it follows back the path described in each NDN node. This is what the authors calls "bread crumbs".
3) Forwarding Information Base (FIB): it stores information about potential location of content matching, forwarding I-packet to the data source. The NDN FIB is very similar to IP FIB table, but due to the NDN philosophy, it is not limited by spanning tree, it can use the advantage of multiple face's sources.

A longest-match lookup is done on its *Content Name* field every time an interest packet arrives on any face. There is a

 ISBN: 978-1-61208-398-8

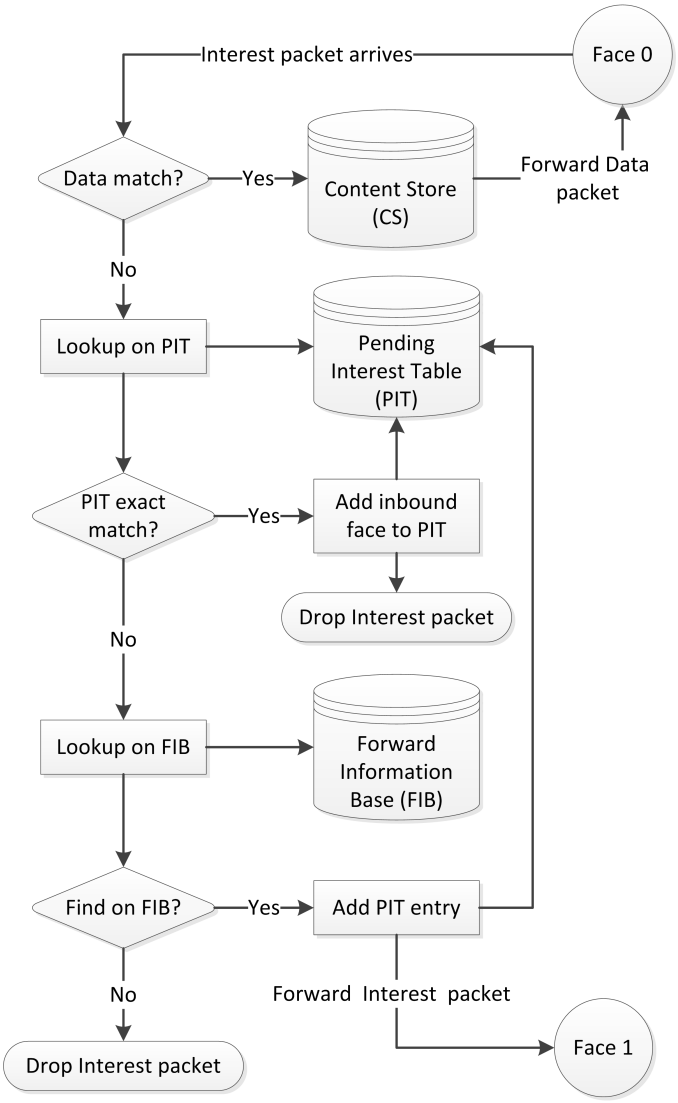


Figure 1. NDN basic operation

structure index to sort a precedence of search, first *Content Store*, next PIT and then, FIB.

If there is a data matching in Content Store, with the same prefix name, the content is sent back through arrival face. Otherwise if there are no data matching, it made a lookup in the PIT table. If there is a prefix match in PIT, the face is added in the Request Face List, and the Interest packet is dropped.

Otherwise, a lookup is done on FIB table. If there is a matching entry, next it includes an entry on PIT table, indicating the face where the pending interest was done. Then, the Interest packet is forwarded to the potential face pointed by FIB table.

If there is any matching entry, the Interest packet is dropped. This process is shown in Figure 1.

The data packet has a simple mechanism; it does not need to be routed. It just needs to track the path created by Interest packets in each NDN node. The path is traced, through PIT entries chain, until the origin request. It follows the "bread crumbs".

When a data packet arrives in NDN node, it is done a lookup by prefix name in Content Store. If there is an entry matching, it means a duplicate content, and then, this data packet is dropped.

Another packet discards occurs when the Data packet not match any entry in PIT, it means that this Data was not required before, it did not receive any Interested.

However, if there is a matching entry in PIT, the data were required by a face. The Data packet is authenticated and added to Content Store. Then, it is created a list with all faces that requested this Interest, and the Data packet is sent to each face in the list.

The treatment of Interest and Data packets, allows to retrieve content only by content name, that is simple and robust. Moreover, NDN is not limited by loops in layer 2; therefore, NDN take advantages of multiple face's sources, processing parallel requests. The NDN hop-by-hop information forwarding does not need to link layer 3 addresses to layer 2 identities, like IP and MAC address. Each NDN node can use information from the request packet. The request time and rate are able to measure the best way finding an interest.

Amendments initiated by the NDN model reflects it is more suitable to content distribution. According to evaluation by Jacobson's [6], showing that, comparing content dissemination performance between TCP/IP networks and NDN, the NDN approach does not increase the traffic according to the number of the users. Basically, for a unique client, TCP/IP was better than NDN. However, while the number of clients increase the total download time increases linearly proportional to the number of client. Otherwise, NDN network maintains download time constant with client number increases.

In NDN architecture, contents may be cached at intermediate nodes along the path from content providers to content users. This strategy, called *on-path cache*, provides contents near to users, reducing the bandwidth and the retrieval time. However, some works demonstrated that this strategy is not efficient because it may imply in a high content replication in nodes.

Another approach, called *off-path cache*, can reduce duplication maintaining the overall hit rate. This approach consists in three strategies: (1) caching only the most popular contents; (2) choose the best cache to push the content improving the hit ratio; and (3) redirect user's requests to the best cache.

The NDN network model shows a simple and flexible proposal, desired to dynamic nature of current Internet. However, it is very difficult to deploy this model purely, due to many open issues, like security and point-to-point applications like VoIP, video stream, e-mail, games, and others. Then, it is important to validate the NDN network model to solid web applications.

Thus, the best way to use NDN is over an IP infrastructure, to disseminate popular content and relevant size. The reduction in the download time for users, and better accuracy of investments, required by producers of content for allocation and content distribution, are benefits expected.

CCNx is the official implementation of NDN, which is currently under development. To be compatible with the existing architecture, CCNx builds an IP overlay to transport Interest and Data packets. The current version of CCNx is 0.8.2, and it supports Linux and Android platform [14].

 ISBN: 978-1-61208-398-8

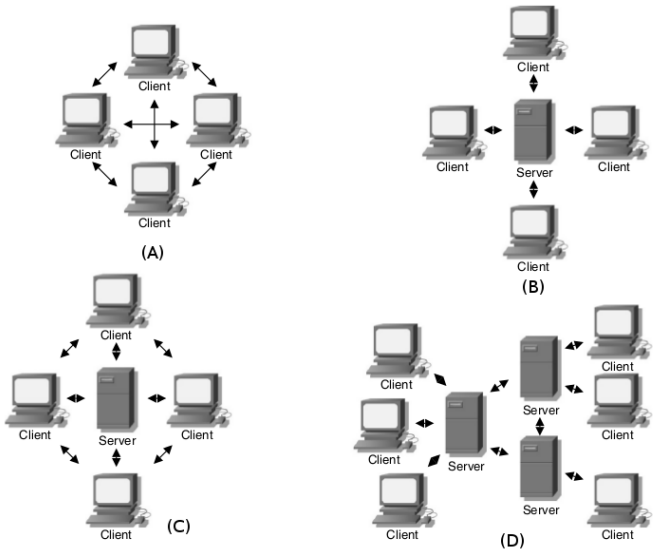


Figure 2. Game network architectures [18]

IV. Online Games Architecture

The most-recent game market research [15] have shown increased from 41 billion US dollars in 2007 to 68 billion in 2012, an increase of 10% per annum. In order to attend the market demand, was designed the Digital Distribution, where the content is delivered on a digital format, dispensing a physical media. This distribution scheme became more viable from 2000, accompanying the growth and evolution of the telecommunication network's and Internet bandwidth increase.

Most of the big gaming companies developed their own platform of distribution content. The games' distribution platforms were deployed, such as Steam, Origin, Live, PSN [16]. The basic idea is to provide a central service to storing contents in a digital format. Moreover, the platform also delivers other contents related to games, like movies and soundtrack. To provide QoS in a content distribution system, it is required a huge investment on network infrastructure. Steam has 8 millions costumers [17], and 13.5 PB of content per week in USA, representing 21.2% of global Internet traffic [2].

The games' market has popularity and a huge content dissemination, giving many opportunities to create new distribution architecture, e.g., using NDN networks.

A. Game Communication Architectures

Before presenting the NDNGame architecture proposal, it is necessary to know how the legacy network game infrastructure works. This section presents an overview of online game's evolution and some interesting issues used in this proposal. The communication model about network games is not different from the legacy network to distributed applications, i.e., peer-to-peer and client-server. However, each architecture produces little relevant modifications to our proposal. We can see an overview in Figure 2.

B. Peer-to-peer

In 1993, DOOM was the first First-Person Shooter (FPS) game with multiplayer cooperation mode up to four players. This game used peer-to-peer model in LAN over IPX protocol with broadcast transmission. Each player runs a game instance locally, and it should send messages to other players in a decentralized way [18].

The main challenges in the peer-to-peer model are related to synchronization. The game should be completely deterministic; each machine should have to execute the same set of instruction in a specific time interval, independent of network behavior. To guarantee this feature, it is necessary to wait a certain time until all players receive the messages in order to update the game state. Then, the game latency is equal for all players, i.e., the biggest delay of all players. This model has been used for Real-Time Strategy (RTS) games [19].

C. Client-server

The peer-to-peer game model works well over the LAN, but not on the Internet. The nature of the global Internet supposes that some users could have low bandwidth links producing long delays. As the game latency is the largest delay of all players, if one is connected in a high delay link, the game experience became bad for all players.

To overcome this problem, in 1996, Quake 3 was released using a client-server architecture [20]. In Quake 3 all players (client) send control messages and update messages only to one machine, a centralized communication server. The clients send to the server all necessary information to process the game state. The server receives and processes the next game state, and it sends a response to all players in order to update the client local state [19].

However, the client machine has just an approximation of actual game state, working like a "dumb terminal". Then, it is not necessary to guarantee deterministic game state in all player's machine; the real game state occurs only on the server. The game Quality of Experience (QoE) is directly related to latency between client-server. As the delay, inside network backbone is significantly lower compared to client bandwidth, the game experience in client-server architecture is better than peer-to-peer architecture. This model is adopted by most online games companies [21].

The evolution of the client-server model is the server-pool [21]. In this model, there is a server pool, in peer-to-peer or client-server architecture, connected to local servers near to the clients. A client can connect to the server pool, through the local server. Server pool increases the architecture complexity, but provides better scalability and game experience for users.

The last model is a combination of client-server with peer-to-peer model [18]. A hybrid network can provide the best from both worlds. The game control plane works on the central server, like traditional client-server. However, the clients are able to connect direct to exchange information and process the game state. This approach is used in VoIP and message chats.

V. The NDNGame Architecture

According to the previous section, the most-used model by game companies is the client-server. Problems like scalability and unique point of failure are recurrent on this approach [22]. For example, in a new game release or Downloadable Content (DLC), it is normal a huge increase on server's load. And this overload might cause failures and service interruption [23].

To reduce the failure risk, the game companies invest in network infrastructure like virtual machine allocations and cloud computing architecture.

 ISBN: 978-1-61208-398-8

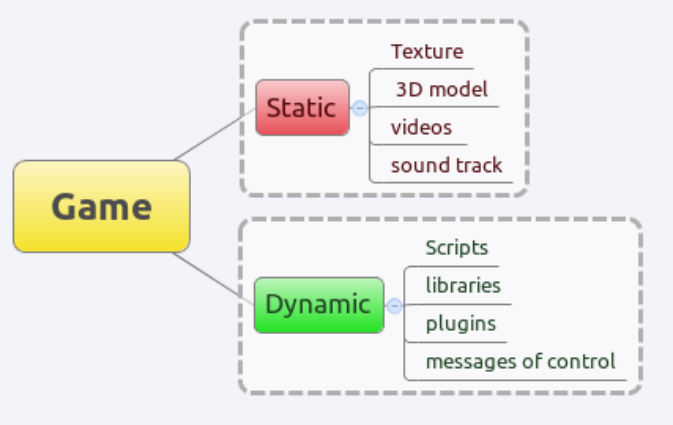


Figure 3. Abstract content

Due to these issues, our approach uses a hybrid network model. A unique local server can support a limited client number. However, if there is content sharing among clients, the local server will be able to support a large client number, i.e., more clients with same infrastructure.

The publisher releases the content on the local server to disseminate it to the first cache. Next, each client will be able to release same content on NDN network. Then, the content availability is directly proportional to demand. This approach is scalable and there is not any central point of failure.

A. *Content Classification*

The basic idea is to classify games content in two types: static and dynamic content. Figure 3 shows the content classification.

A multimedia game is composed by files related to environment and characters, as texture used to building maps and characters. Computer graphic scene and sound track also comprise this type of packet. This content represents most part of bandwidth, and it does not change from client to client. Storing and renderization are made locally using the player device. So, this content is declared as static, viable to share among all clients.

On the other hand, logical parts as scripts, libraries, plugins, control messages, have a dynamic behavior, change over time. They are dependent of hardware and client Operational System (OS), moreover, requires lower bandwidth.

Dynamic packets are extremely sensible to network latency. The response time of dynamic content is very important for QoE. According to Chen [24], the game-play time is reduced when network latency is increased. Thus, the best way to deliver this content is on traditional client-server networks based on TCP/IP, or even better, UDP/IP protocol.

The multimedia packets have other necessities. This content causes a great impact on server load, due to larger files. We believe the best strategy to work this content is to divide the responsibility among clients. The client community can share multimedia packets for the same game, reducing traffic on local servers.

In the architecture overview, we have a traditional client-server network, working on TPC/IP as base for a NDN overlay network. We basically split content traffic between static and dynamic, and forward it to the network layer which best attends to player's need. We can see an overview in Figure 4.

When a client purchases a new digital game, it downloads the game's dynamic pack with the authenticated files. The entire transaction is done by the traditional IP network. After that, the publisher sends a list of static content, which can attend to requests from the new recently deployed game. Cache's networks are composed by the local server and the clients who have the game or just part of it.

A client is also an NDN node, possessing a FIB. This database will be fed by the local server, updating the new cache availability. Thus, it there is not the necessity to broadcast the interest packets, it is only needed to forward them to caches, which have the requested content. This will help to avoid packet's flood, redirecting the data flow on the network. After send the Interest packet, clients wait for desired data while it shares the content already downloaded.

The NDNGame proposal provides some advantages:

- Low complexity, our proposal works on application's layer; it is not necessary to modify the basic structure of IP packets, neither NDN network core. In this manner, there is a great chance of success to deployment this model in the current game's market.
- Cost reduction: the content sharing by the users, provides a bandwidth reduction on content local servers. Thus, it is possible to attend more clients and to save infrastructure investments.
- Availability and Scalability: the content network infrastructures work on demand, thus it does not degrade if the number of player increases. Moreover, the availability also increases as the demand increases; more clients mean more available caches.

VI. Proposal Evaluation

To evaluate the NDNGame proposal was built two scenarios: (1) a legacy network with ten IP switches, showed in Figure 5, and (2) a hybrid NDN and IP network ten NDN switch with an IP routing table, showed in Figure 6. The scenario is a topology with ten switches in line. In each scenario, we consider one gaming server and ten players. The game traffic used was the Quake 3. The Quake 3 game application was chosen in order to compare against some related work, which uses it as reference game application. We evaluate the proposed NDNGame architecture against a legacy IP network.

The evaluation environment was based on a virtualized network using Mininet [7]. The Mininet system permits the specification of a network interconnecting "virtualized" devices. Each network device, hosts, switches and controller are virtualized and communicate via Mininet. A Python script is used to create the topology in Mininet, and the traffic flows controls are made by the OpenFlow controller. Therefore, the test environment implements and performs the actual protocol stacks that communicate with each other virtually. The Mininet environment allows the execution of real protocols in a virtual network.

The experiment was built over VMware Workstation 10. All tests run on a Dell PowerEdge R-620 with two Xeon processors and 64 GB RAM. To ensure the reliability of results, the workload in the server is maintained below 80% of processor capacity.

 ISBN: 978-1-61208-398-8

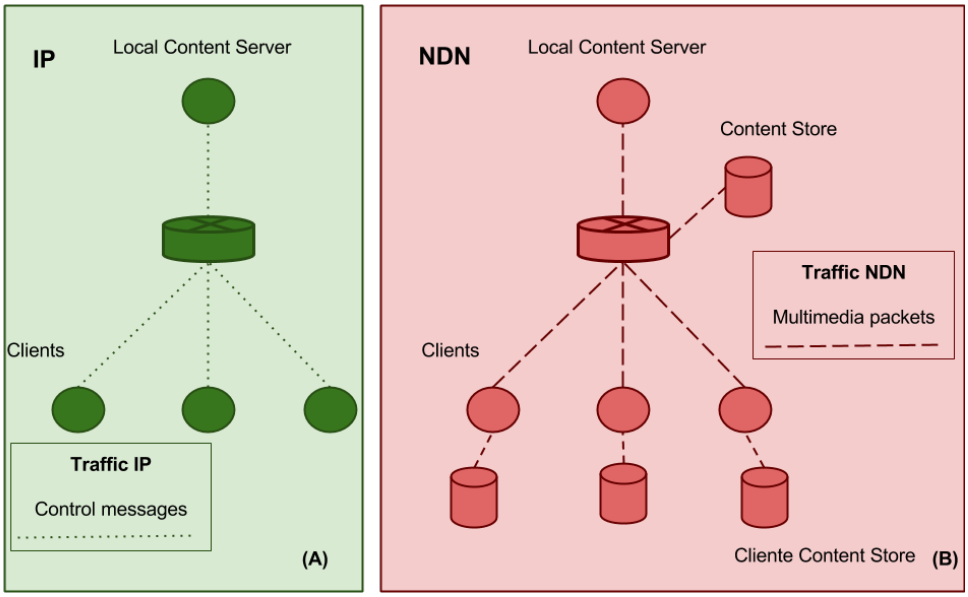


Figure 4. NDNGame architecture: (A) IP network and (B) NDN network.

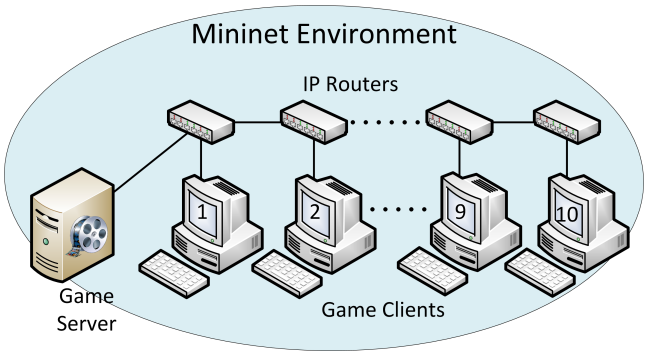


Figure 5. Legacy network evaluation scenario

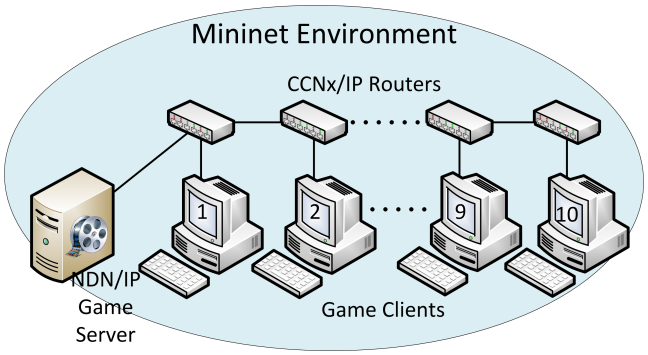


Figure 6. NDNGame evaluation scenario

The NDNGame prototype was built over Mini-CCNx environment [25], a CCNx implementation in Mininet. As CCNx is a CCN overlay over IP, it is possible to implement a hybrid network. The Quake 3 server forwards all static content packets to UDP port 9695. So, all 9695 UDP packets are treated as NDN packet across the overlay network. The Quake 3 game traffic was generated by D-ITG software [26] and a proxy inside servers classify static and dynamic content.

In initial evaluation, the results show an improvement in the delay of game content distribution when the number of users increase and a reduction on network and server load.

VII. Conclusion and Future Work

The NDN model works well to distribute massive amounts of static content. However, point-to-point application is not suitable to this model, due to NDN retrieve content by name instead an address. This problem makes it difficult for a pure NDN network deployment.

To overcome this barrier, this paper proposed a new hybrid network game architecture where NDN networks could be used to improve online game's distribution network. It is done applying NDN for content dissemination and point-to-point IP communication to transmit control messages. Our approach is simple and provides scalability and cost reduction maintaining the user experience.

A prototype was deployed, and the initial results showed that the proposal could reduce the network delay and reduce the network and servers' load, providing a better user experience.

As future work, we intend to evaluate the proposal in more diverse scenario and traffic workload. Another important issue

 ISBN: 978-1-61208-398-8

is to analyze the impact of dual protocol stack, IP and NDN, in a unique switch.

REFERENCES

[1] J. S. Turner and D. E. Taylor, "Diversifying the internet," in Global Telecommunications Conference, 2005. GLOBECOM'05. IEEE, vol. 2. IEEE, 2005, pp. 760–766.

[2] Steam, "Steam download stats," 2014, URL: http://store.steampowered.com/stats/content [retrieved: March, 2015].

[3] Theesa, "Essential facts about the computer and video game industry," 2014, URL: http://www.theesa.com/facts/pdfs/esa_ef_2014.pdf [retrieved: March, 2015].

[4] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the Future, vol. 2007, 2012, pp. 1–16.

[5] A. Feldmann, "Internet clean-slate design: what and why?" ACM SIGCOMM Computer Communication Review, vol. 37, no. 3, 2007, pp. 59–64.

[6] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in Proceedings of the 5th international conference on Emerging networking experiments and technologies. ACM, 2009, pp. 1–12.

[7] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: rapid prototyping for software-defined networks," in Proceedings of the Ninth ACM SIGCOMM Workshop on Hot Topics in Networks, ser. Hotnets '10. New York, NY, USA: ACM, 2010, pp. 19:1–19:6.

[8] A. Bharambe, J. R. Douceur, J. R. Lorch, T. Moscibroda, J. Pang, S. Seshan, and X. Zhuang, "Donnybrook: enabling large-scale, high-speed, peer-to-peer games," ACM SIGCOMM Computer Communication Review, vol. 38, no. 4, 2008, pp. 389–400.

[9] V. Jacobson, M. Mosko, D. Smetters, and J. Garcia-Luna-Aceves, "Content-centric networking," Whitepaper, Palo Alto Research Center, 2007, pp. 2–4.

[10] V. Jacobson, D. K. Smetters, N. H. Briggs, and et al., "Voccn: voice-over content-centric networks," in Proceedings of the 2009 workshop on Re-architecting the internet. ACM, 2009, pp. 1–6.

[11] J. Chen, M. Arumaithurai, X. Fu, and K. Ramakrishnan, "G-copss: a content centric communication infrastructure for gaming applications," in Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on. IEEE, 2012, pp. 355–365.

[12] J. Chen, M. Arumaithurai, L. Jiao, X. Fu, and K. Ramakrishnan, "Copss: An efficient content oriented publish/subscribe system," in Architectures for Networking and Communications Systems (ANCS), 2011 Seventh ACM/IEEE Symposium on. IEEE, 2011, pp. 99–110.

[13] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, and et al., "A data-oriented (and beyond) network architecture," ACM SIGCOMM Computer Communication Review, vol. 37, no. 4, 2007, pp. 181–192.

[14] "CCNx Project," 2014, URL: http://www.ccnx.org/ [retrieved: March, 2015].

[15] F. Caron, "Gaming expected to be a 68 billion business by 2012," 2014, URL: http://www.arstechnica.com/gaming [retrieved: March, 2015].

[16] Wikipedia, "Digital distribution in video games," 2014, URL: http://en.wikipedia.org/wiki/Digital_distribution_in_video_games [retrieved: March, 2015].

[17] ——, "Steam software," 2014, URL: http://en.wikipedia.org/wiki/Steam_28software29 [retrieved: March, 2015].

[18] G. Armitage, M. Claypool, and P. Branch, Networking and online games: understanding and engineering multiplayer Internet games. John Wiley & Sons, 2006.

[19] GafferonGames, "What every programmer needs to know about game networking," 2014, URL: http://gafferongames.com/networking-for-game-programmers/what-every-programmer-needs-to-know-about-game-networking/ [retrieved: March, 2015].

[20] F. Sanglard, "Quake 3 source code review: Network model," 2014, URL: http://fabiensanglard.net/quake3/network.php [retrieved: March, 2015].

[21] J. Smed, T. Kaukoranta, and H. Hakonen, "Aspects of networking in multiplayer computer games," The Electronic Library, vol. 20, no. 2, 2002, pp. 87–97.

[22] A. Tanenbaum and M. Van Steen, Distributed systems. Pearson Prentice Hall, 2007.

[23] R. Appleton, "Titanfall's pc servers are overloaded," URL:http://www.gamerheadlines.com/2014/03 [retrieved: March, 2015].

[24] K.-T. Chen, P. Huang, and C.-L. Lei, "How sensitive are online gamers to network quality?" Communications of the ACM, vol. 49, no. 11, 2006, pp. 34–38.

[25] C. Cabral, C. E. Rothenberg, and M. F. Magalhães, "Mini-CCNx: fast prototyping for named data networking," in Proceedings of the 3rd ACM SIGCOMM workshop on Information-centric networking. ACM, 2013, pp. 33–34.

[26] S. Avallone, S. Guadagno, D. Emma, A. Pescape, and G. Ventre, "D-itg distributed internet traffic generator," in Quantitative Evaluation of Systems, 2004. QEST 2004. Proceedings. First International Conference on the. IEEE, 2004, pp. 316–317.

 ISBN: 978-1-61208-398-8

Revisiting Virtual Private Network Service at Carrier Networks: Taking Advantage of Software Defined Networking and Network Function Virtualisation

Luiz Cláudio Theodoro, Pedro Macedo Leite
Hélvio Pereira de Freitas, Adailson Carlos Souza Passos,
and João Henrique de Souza Pereira

Innovation, Research and Development
Algar Telecom
Uberlândia, MG, Brazil
Email: lclaudio@algartelecom.com.br,
pedrol@algartelecom.com.br, helvio@algartelecom.com.br,
adailson@algartelecom.com.br, joaohs@algartelecom.com.br

Flávio de Oliveira Silva, Pedro Frosi Rosa,
João Henrique de Souza Pereira
and Alexandre Cardoso

Federal University of Uberlândia
Uberlândia, MG, Brazil
Email: flavio@ufu.br, pfrosi@ufu.br,
joaohs@ufu.br, and alexandre@ufu.br

***Abstract*—A service commonly offered by telecom operators is the Virtual Private Network (VPN) that allows the interconnection of corporate networks in different geographic localities. To deploy this service, an operator, usually, installs in customer's premises some equipments, such as a router and/or a switch. This deployment increases cost and the complexity of the service. By taking advantage of Software Defined Networking (SDN) and Network Function Virtualisation (NFV), this work presents and details a VPN service architecture that decreases Capital Expense (CAPEX) and Operating Expense (OPEX) and it is open to new innovative service offerings, such as: Quality of Service (QoS) policies, Network Address Translation (NAT) and a multi-homing function. This work also contributes to this research area by going further in the description of a real use case at a carrier network.**

Keywords*—*Software Defined Networking; Network Function Virtualisation; VPN; QoS; NAT.

I. INTRODUCTION

Large computer networks are difficult to manage and are not simple at all in its structure for the fact they involve several equipments, such as switches, routers, modems, servers and others. On these equipments, firewalls, Network Address Translation (NATs), servers load balancers and intrusion-detection systems are configured. The industry has been delivering to the market an infinite of equipment which work in a complex way with a distributed control software closed and proprietary. Such software implements network protocols exhaustively tested during many years that evolved a lot in terms of standardisation and interoperability and generate a big industry based on IP services.

In a traditional scenario, IP Services are provided by a Customer Premises Equipments (CPE), which are the point of contact with the customer facility. All the configuration related to user services resides on the CPE, and in case of any change it must be remotely accessed and updated accordingly. If there is a hardware failure a field technician must be in place in order to substitute the CPE. In this case, the downtime is considerable. An alternative would be to implement a resilient structure, with spare parts and on-line redundancy, thus increasing costs and complexity.

From the operator perspective, this also implies additional costs with technician displacement, keeping spare parts (sometimes from different CPE vendors) and maintenance contracts with third parties.

Normally, administrators configure individual network devices using configuration interfaces that vary among suppliers and even among different products from the same supplier. Though some network management tools offer a central management entity for the network configuration these systems continue to operate at individual configuration protocols, mechanisms and interface levels. Such an operation mode hampers innovation, increases complexity and encumbers companies with high investments and high network operation costs [1][2].

As a proposal to change the way the networks are projected and managed there was Software Defined Networking (SDN) abstractions that separate the control plane (that decides how to control the traffic) from the data plane (that forwards data according to the control plane decisions). By shifting control plane functions to a central place could result in a more proactive management besides optimizing CPE maintenance costs and potentially reducing displacements to customer facilities.

Another important feature SDN has is to use a well-defined Application Programming Interface (API) as the OpenFlow [1], in order to have direct control on the data plane elements; its acceptance has been increasing within the industry and it was the greatest responsible for taking the SDN from the academy to the telecom marketplace. Several commercial switches can already support OpenFlow protocol and several control platforms have been launched [3].

Recently, a new initiative, called Network Functions Virtualisation (NVF), helped to launch virtualisation existent concepts to consolidate network equipment with specific functions in servers with high volume. Switches and storage that can be allocated on the network nodes, on datacentres or on the final user equipment [4][5].

 ISBN: 978-1-61208-398-8

These two technologies have already crossed the labs barriers and have entered for once on the producer's road-map. Although, SDN has benefited some initial practical successes and it certainly offers necessary technologies to support network virtualisation, lots of work is needed in either to improve the existent infrastructure or to explore SDN's potential in order to solve the problems from a much wider perspectives of use cases. This work is part of such a movement and it points to a real scenario that can be implemented to collaborate on these technologies opportunities and challenges [2].

Both concepts and solutions can be potentially combined in a way to obtain more added value to companies that provide the service and to their customers.

This work revisits the VPN service, a common service, offered by telecom operators that allows the interconnection of networks in different localities. The new service architecture reduces CAPEX and OPEX and also offers innovative functions that can empower the customer by giving the ability to explore new service funcionalities such as different QoS policies, NAT and multi-homing.

The remainder of this paper is structured as follows: Section II describes related work regarding NFV and SDN. Section III presents the current deployment of a VPN service by telecom operators. Section IV revisits the VPN service deployment based on SDN and NFV. Section V describes some innovative service offerings built on top this new service architecture, and finally, in Section VI, we present some concluding remarks and future work.

II. Related Work

Many companies are on the search for solutions to improve their services, to reduce costs and to increase innovation perspectives. Therefore, NFV became an excellent option and thus a great number of researches from allover the world are working for its evolution. In parallel, SDN, as a solution to control the network has been increasing and it can act in consonance with the proposed NFV to give better conditions for implementing new solutions. The NFV is highly complementary to the SDN, but it is not dependent of this one (and vice and versa). SDN aims at dissociating the control plane from the data plane and it was projected as an architecture that uses a centralised control plane in order to ensure better scalability and agility for large networks. Several papers proposed techniques to overcome scalability limitation and they are being effectively implemented by companies looking for obtaining great benefits [4][6][7].

On the other hand, there is a motivation to face NFV and SDN's technical and business challenges with firm intention to clarify functions and interactions from several kinds of commercial entities that act on the market with such technologies. An example can be seen in a use case set shown by European Telecommunications Standard Institute (ETSI) [8].

As the industry closely keeps up with the evolution of these technologies , many have already addressed on the problem and they contribute offering subsides for a better understanding of possibilities from these innovations. From the moment the experiments left the academy and migrated to Wide Area Network (WAN); many use cases and challenges were presented. In this process, network operators were mobilised to disseminate the network virtualisation practice using the virtualisation efficient concepts already intensely adopted by Information Technology (IT) areas and hardware commoditisation for WAN application [9].

Also, considering ETSI's orientations [8] potential use cases were assimilated for the NFV, previously described. They affirm that SDN and NFV are two separated technologies; but, they overlap themselves, each one using the potential of the other one. The orientations mainly point out some challenges to be overcome for these technologies consolidations either on the networks' organisation side or on the IT side, specifically from the cloud.

There are many barriers for the SDN/NFV adoption, for example, the lack of patterns in some areas, the fact the applications have not been projected to be processed in the cloud, the need for interoperability with the infrastructure and legacy systems and others issues but nevertheless, there are high expectations regarding their use [9].

One of the first implementations for functional NFV's concept by means of the forwarding virtualisation function through an OpenFlow network deals with the current IPv4 and IPv6 coexistence and the possibilities brought for the arena by enabled OpenFlow infrastructure. The routing virtualised protocol project is described allowing a simple management and avoiding signalling message overload at the level of the control plane and also avoiding different scenarios considered in order to validate the virtualised function [7].

To help the researchers to accelerate the proposals, in 2014, UNIFY [10] project was born, that aim to open up the potential of virtualisation and automation across the whole networking and cloud infrastructure developing an automated, dynamic service creation platform; thus, leveraging a fine-granular service chaining architecture. This new solution has motivated researchers who will use soon a global orchestrator with optimisation algorithms to ensure optimal placement of elementary service components across the infrastructure. UNIFY launches the NIB (Network Information Base) concept that captures the network aspects and mounts a map of network and processing resources as well as their current state. Interacting with NIB, there are elements responsible for dynamically orchestrating network functions and resources.

Many aspects from the related work are considered on the proposition of the current use case in this paper. Other practises are revealed in other publications as the effort to implement SDN/NFV on Mobile Backhaul Networks. As a conclusion, platforms capable of enabling the SDN/NFV service that show the fact that uniting both these technologies is a current demand [1][2].

III. VPN Current Deployed Approach

Let us take an example a service, such as VPN, that allows to join corporate networks in different geographic localities.

The VPN service requires a very complex structure involving many equipments like switches routers, Firewalls, etc. In a typical VPN L3 scenario (RFC2547 - BGP/MPLS VPNs), a router is installed at customer facility and connected to the nearest border router or Provider Edg (PE) using a Time Division Multiplex (TDM). At the CPE side, separated

 ISBN: 978-1-61208-398-8

routing instances, named Virtual Routing Function (VRF) Lite (which is a logical way of segregating network traffic) handle voice and data traffic; QoS is applied accordingly. In a Metro Ethernet scenario, voice and data are carried using different Virtual Local Area Networks (VLANs) over an Ethernet link, using appropriate QoS marking. At the PE, each VPN has its own VRF (a separated routing table) which handles all the different user traffic. CPE management is accomplished using special policies that allow Management Servers to ping and get SNMP statistics without address conflicts [11].

Some customers require different CPE for voice and data whereas others require high availability or more flexible bandwidth management. A customer activation process is very complex due to the necessity of taking many steps for the effective service implantation: acquisition, installation, configuration and elements operation; also, the operator's team is trained in different owned hardware. The most bureaucratic and delayed activity is the Customer Premises Equipment (CPE) installation at the client's location. This last installation demands operator's time, client's time, adaptation of the structure that will receive the equipment (energy, temperature, cabling, etc.). An infrastructure example to supply this service is shown in Figure 1.

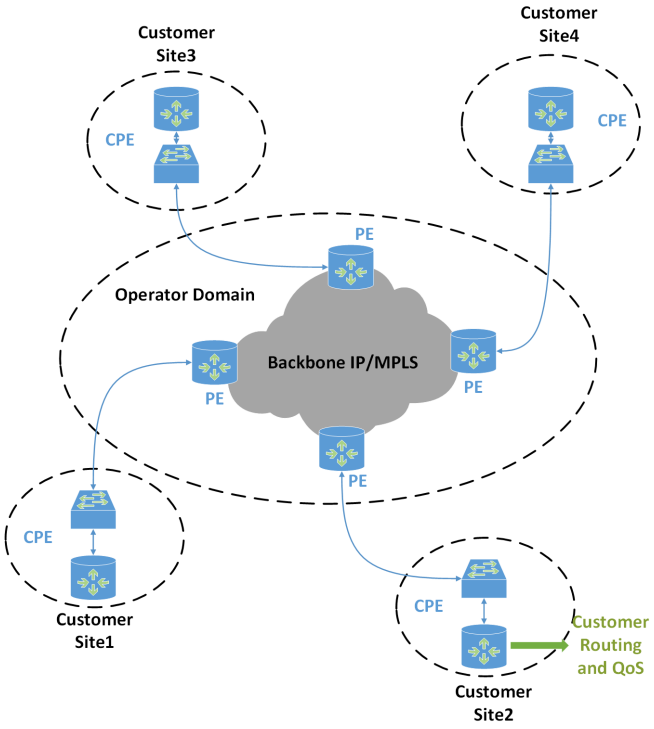


Figure 1. Current Infrastructure of a VPN Service

The image provides a superficial overview and without deepening in the elements we can identify a backbone network on MPLS, a Metro-Ethernet network and the client's CPE. The Multiprotocol Label Switching (MPLS) network, according to RFC 3031, is a framework that allows traffic flow forwarding and its efficient commutation through the network; this means, it is a switch entanglement controlled by MPLS that can supply a broad amount of traffic satisfactory. When it comes to the Metro-Ethernet network, it is a set of switches interconnecting in layer 2; so, by using only the Ethernet protocol [12].

Therefore, in order to have a dataflow from a branch of a certain client for another branch on a geographically separated locality, it is necessary the setting of several switches with different technologies. If in one of these switches the setting is incorrect, the service certainly does not work. The CPE is the equipment located at the client, providing a specific service, which, in this case, is the VPN tunnel and proper client's network routing. Beside both these functions, many others can be added, such as firewall, proxy, WEB server, etc.

The CPE has to be set at the client's location and it also requires maintenance. In case of this equipment's bad functioning, it directly affects the VPN. When VPN does not work, the process inside the operator is to change the CPE, and then, to reconfigure settings. This process has been happening routinely in many operators around the world and a big cost taken on by the operators are the activities in the customer's site to install, configure and mantains the CPEs. The value calculated for this operations is very high and compromises much of the revenue reducing considerably the profit of the carriers and consequently burdening the service value for the customers.

On other hand, due to the increasing costs of the TDM infrastructure, serial links have been gradually replaced by Metro Ethernet access in the last mile. Again, due to licensing costs, an additional switch performing L2 functions, which also provides path redundancy, has been added between the CPE router and the access ring, as shown in Figure 1.

IV. VPN Service Based on SDN and NFV

Taking advantage of SDN and NFV, this section presents a VPN service architecture which reduces service CAPEX and OPEX, and offers innovative functionalities to customers. Considering the most common used VPN service (as described in Section III), this work assumes the SDN/NFV implementation for a VPN service deployed on top of a ring Metro-Ethernet network.

By using virtualisation techniques, CPE functions were moved to the cloud, thus leaving a simpler and cheaper equipment at the customer facility. Handling user traffic would not be constrained by CPU and memory resources, since they can be added on demand, thus leaving room for more innovative services and a better response to the changing and unbalanced traffic (traffic optimisation).

As depicted in Figure 2, at the customer facility, the access switch is replaced by an OpenFlow capable switch. The physical router is removed and its function is virtualised at the *Cloud Router*.

Located at the telecom data centre, the *Cloud Router* is a new network entity. Each physical router is deployed as a virtual machine that runs the Quagga Software Route Suite [13] that provides implementations of Open Shortest Path First (OSPF), Routing Information Protoco (RIP), and Border Gateway Protocol 4 (BGP-4). Each virtualised router offers two Representational State Transfer (REST) based APIs. One is used by Operations Support Systems (OSS) in order to configure the service and its functionalities. This configuration can be done by the operator or by the customer.

 ISBN: 978-1-61208-398-8

The other REST API is used by the *SDN Control Layer* in order to query the *Cloud Router* about the routes and service updates. The *SDN Control Layer* is logically centralised at the operator backbone and is responsible to configure the OpenFlow capable switch accordingly.

Each instance of the *Cloud Router* will handle route exchanges with the PE, eg., RIP, OSPF and also will feed the *SDN Control Layer* with routing updates. The *SDN Control Layer* will handle QoS and NAT accordingly, applying match and action rules to each OpenFlow switch, as depicted in Figure 2.

Customer traffic is carried in a private VLAN tag which maps to a sub-interface on the PE side. This sub-interface belongs to the customer Virtual Routing and Forwarding (VRF) which stores all the routes from customer VPN remote sites. In order to exchange route updates the *SDN Control Layer* has a connection the the customer VRF. This can pose a problem if a single controller is used for different customers where Internet Protocol (IP) addresses overlap. However, this can be overcome by using a reserved IP range for management purposes, thus mitigating IP conflicts, and import/export policies in order to make management servers and CPEs visible to each other.

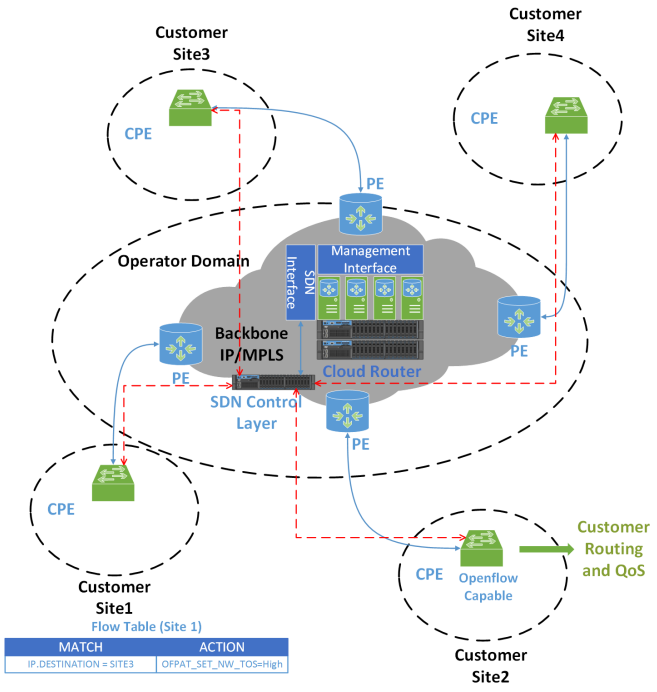


Figure 2. SDN/NFV VPN Architecture

Effectively, the customers stop having a router and can have a smaller device, which can be cheaper allowing them to reach the switch. The final product cost is reduced for the client, the operation becomes simpler, router installation and maintenance is eliminated and this brings huge benefits for the company and the final user.

In the traditional switch, we have the Control Plane that has the entire network intelligence and it owns each supplier. Since Data Plane will not change for the OpenFlow switch; only the smart part of the traditional switch will be centralised and the available protocols on the switch will give place to the OpenFlow secure channel keeping the Data Plane unaltered.

At the real world, a radical implementation is hardly used, thus, the most common way will be the coexistence of traditional protocols with the OpenFlow part on a hybrid composition. This allows this technology to treat the network's resilience, providing security and redundancy, as a result. The switch can support the OpenFlow, and, at the same time, the traditional protocols so the ring resilience to commutating when necessary can be done with the switch's traditional protocols, Spanning Tree, for example. Effectively, the traditional protocol can be used for commuting the ring resilience, and, at the same time, enabling the switch's port to implement the SDN.

To give a detailed vision, the topology containing the interfaces and modules is given below. The client's CPE (router) is virtualised at the central environment; in practice, it is implemented by a Virtual Machine (VM). This VM performs the functions originally executed for the proprietary hardware (routing, QoS, etc.). The details of the functional model that composes the Virtual CPE can be seen in the Figure 3, including the connections.

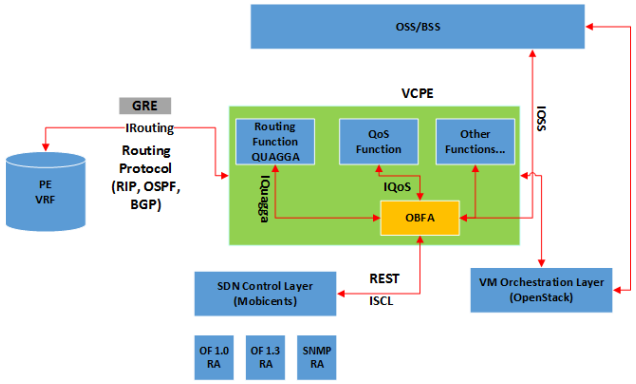


Figure 3. VPN Interfaces diagram

The Virtual Customer Router Function (VCRF) Module interacts with router PE from operator, by using the IRouting Interface and is responsible for exchanging of routes update messages. As the VCPE can be located in some network point not directly connected, a way is required to connect the VCRF to the appropriate PE, so that the routing protocol can establish its adjacency. In this case, the Generic Routing Encapsulation (GRE) is used. The VCRF interacts with the Orchestration and Business Function Aggregation (OBFA) Module trough IQUAGGA interface providing updated information of the routing table. With the respect to QoS requirements related to the client traffic, the module Virtual QoS Function (VQoSF) connected to OFBA Module is used.

The Module OBFA, based on the information received by the VCRF, VQoSF, and IOSS (Interface OSS) creates the suitable flows and communicates to SDN Controller using REST or Interface SDN Control Language (ISCL). Thus, the SDN Controller sends the configuration to the client's switch. The communication with OpenFlows switches is made by the Resource Adaptor (RA) that implements the version of the

 ISBN: 978-1-61208-398-8

protocol suported by customer CPE.

V. USE CASES

The VPN service described in Section IV enables innovative service offerings that can be provided to customers, by giving them the ability to take the control of the service, optimizing operations and reducing OPEX. This section highlights some use these offerings.

A. Quality of Service (QoS) Policies

By using the Management Interface, the customer can deploy different Quality of Service (QoS) policies. For example, let us assume that the traffic from Site 1 destinated to Site 3 will have a higher priority when compared to the traffic destinated to Site 2 or Site 4. The following steps can be run:

1) The customer indicate this policy at the *Management Interface*;
2) Upon modifications, using a REST API callback mechanism, the *SDN Interface* notifies the *SDN Control Layer*;
3) The *SDN Control Layer* translates these modifications to OpenFlow Actions;
4) CPE switch at Site 1 receives an OFPT_FLOW_MOD where the match field is the Site 3 destination IP with an action OFPAT_SET_NW_TOS in order to set DSCP field with a higher priority considering operator forwarding policies;
5) Traffic from Site 1 to Site 3 will be forwarded with higher priority accordingly to the carrier QoS policies.

The example illustrates how QoS policies can be applied to the service. It is important to notice that several other policies can be further deployed based on customer preferences in a programmable way, as long as the OSS system supports these new functionalities.

B. Network Address Translation(NAT)

Sometimes, it becomes necessary to modify the destination address of the packets coming from a particular site (Site 2 for instance) when destination server is out for maintenance (for example at Site 3). In this simple situation, we want to move traffic to an alternative server (at Site 4), momentarily, until the original server comes up again. Thus, we can use the strategy described below using the Management Interface:

1) The customer indicates the new policy at the *Management Interface*;
2) Upon modifications, using the REST API, the *SDN Interface* notifies the *SDN Control Layer*;
3) *SDN Control Layer* translates this new configuration to OpenFlow actions;
4) CPE switch at Site 2 receives an OFPT_FLOW_MOD where the match field is the Site 3 destination IP with an action OFPAT_SET_NW_DST in order to set IP Destination with the new IP address belonging to Site 4.
5) Traffic from Site 2 to will now be forwarded to new server at Site 4.

C. Multihoming Function

Let us consider a situation where the customer has an alternate router from other operator for backup purposes. This router will be connected to the Openflow switch and it will be used when Site 3 link fails, for example, and its network stops being advertised. In this case, the virtualised CPE from Site 2 will be notified by a Route Update message from the PE routing protocol, e.g., RIP, OSPF. When a Management program receives this event it will ask the SDN Controller from Site 2 (in our example) to set a new rule to deviate the traffic to a pre-defined backup Site (Site 4 for example, which also has an alternate connection), using the following steps:

1) Using a REST API the *Management Interface* communicates with the SDN Control Layer;
2) *SDN Control Layer* translates this context to OpenFlow actions;
3) CPE switch at Site 2 receives an OFPT_FLOW_MOD where the match field is the Site 3 destination IP with an action OFPAT_SET_NW_DST in order to set IP Destination with the alternate IP address belonging to Site 4.
4) Traffic from Site 2 to will now be forwarded to Site 4 using the alternate router.

VI. CONCLUSION AND FUTURE WORK

SDN was adopted by the research community and has had a considerable evolution. Besides this, SDN is present in the roadmap from various manufacturers. Concomitantly, NFV was also widely accepted and gained momentum, especially by applied research that can be exploited by telecom operators. These two promising technologies bring a number of benefits to end users, carrier networks and service providers being essential in innovative scenarios and demonstrating consistent results regarding the feasibility to implement these new solutions.

This work presented an implementation of a VPN service that is currently widely deployed by telecom operators. The service architecture detailed reduces CAPEX and OPEX and may be used to add innovative functions on top of this service, which can empower corporate customers, giving them the control of their service.

The paper also contributed with this research area by presenting to the community a solution deployed at a real telecom operator, thus, fostering NFV and SDN adoption, acting as a blueprint for a VPN service based on these technologies.

Future work will detail results presenting measurements regarding the service and functions in a production environment. Also, new use cases and functionalities will be created and detailed using the presented service architecture as a framework. From this, one can think in future studies aiming at the implementation of new cases and further use of new platforms, such as the recent UNIFY.

ACKNOWLEDGMENT

This work has been partially funded by ALGAR Telecom and the Brazilian agencies: CAPES, CNPq, FAPEMIG and PROPP/UFU.

 ISBN: 978-1-61208-398-8

REFERENCES

[1] N. McKeown et al., "Openflow: enabling innovation in campus networks," ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, 2008, pp. 69–74.

[2] N. Feamster, J. Rexford, and E. Zegura, "The road to SDN: An intellectual history of programmable networks," SIGCOMM Comput. Commun. Rev., vol. 44, no. 2, Apr. 2014, p. 87–98.

[3] F. Schneider, T. Egawa, S. Schaller, S.-i. Hayano, M. Schöller, and F. Zdarsky, "Standardizations of SDN and its practical implementation," vol. 8, no. 2, Apr. 2014, p. 6.

[4] ETSI, "Network functions virtualisation - an introduction, benefits, enablers, challenges e call for action." Whitepaper, 2012.

[5] D. King and C. Ford., "A critical survey of network functions virtualization (nfv)," 2013.

[6] M. F. Bari et al., "Dynamic controller provisioning in software defined networks," in 2013 9th International Conference on Network and Service Management (CNSM). IEEE, 2013, pp. 18–25.

[7] J. Batalle, J. Ferrer Riera, E. Escalona, and J. Garcia-Espin, "On the implementation of NFV over an OpenFlow infrastructure: Routing function virtualization," in Future Networks and Services (SDN4FNS), 2013 IEEE SDN for, Nov. 2013, pp. 1–6.

[8] E. Group Specification. Network function virtualisation (nfv); use cases. [Online]. Available: http://docbox.etsi.org/ISG/NFV/Open/Published/gs_NFV001v010101p%20-%20Use%20Cases.pdf [retrieved: May, 2013]

[9] S. Perrin and S. Hubbard. Practical Implementation of SDN & NFV in the WAN. [Online]. Available: https://networkbuilders.intel.com/docs/HR-Intel-SDN-WP.pdf [retrieved: May, 2015]

[10] A. Császár et al., "Unifying cloud and carrier network: Eu fp7 project unify," in Utility and Cloud Computing (UCC), 2013 IEEE/ACM 6th International Conference on. IEEE, 2013, pp. 452–457.

[11] E. Rosen and Y. Rekhter. BGP/MPLS VPNs. Published: RFC 2547 (Informational) Obsoleted by RFC 4364. [Online]. Available: http://www.ietf.org/rfc/rfc2547.txt [retrieved: Mar., 2015]

[12] E. Rosen, A. Viswanathan, and R. Callon. Multiprotocol label switching architecture. Published: RFC 3031 (Proposed Standard) Updated by RFCs 6178, 6790. [Online]. Available: http://www.ietf.org/rfc/rfc3031.txt [retrieved: Jan., 2015]

[13] OpenSourceRouting. Quagga software routing suite. [Online]. Available: http://www.nongnu.org/quagga/ [retrieved: Mar., 2015]

 ISBN: 978-1-61208-398-8

Assessing Soft- and Hardware Bottlenecks in PC-based Packet Forwarding Systems

Paul Emmerich, Daniel Raumer, Florian Wohlfart, and Georg Carle

Technische Universität München, Department of Computer Science, Network Architectures and Services
Boltzmannstr. 3, 85748 Garching bei München, Germany
{emmericp|raumer|wohlfart|carle}@net.in.tum.de

***Abstract*—Due to grown capabilities of commodity hardware for packet processing and the high flexibility of software, the use of those systems as alternatives to expensive dedicated networking devices has gained momentum. However, the performance of such PC-based software systems is still low when compared to specialized hardware. In this paper, we analyze the performance of several packet forwarding systems and identify bottlenecks by using profiling techniques. We show that the packet IO in the operating system's network stack is a significant bottleneck and that a six-fold performance increase can be achieved with user space networking frameworks like Intel DPDK.**

Keywords–Linux Router; Intel DPDK; Performance Evaluation; Measurement.

I. Introduction

Software routers and switches which are based on commodity hardware provide high flexibility. The user can combine modules without paying for unnecessary features. Software switches hosted on a single server allow for switching between virtual machines above the physical limit of its 10 GbE network adapters [1]. Additionally, almost any middle box behavior can be added to a software switch. Whole operating systems like the Vyatta Open-Firmware-Router [2] which focus on packet processing on commodity hardware have been created as part of new business models demonstrating the marketability of software routers and switches.

We analyze the performance of Linux IP forwarding, Linux bridge, Open vSwitch (OvS) [3], DPDK L2FWD (a forwarding application based on the user space packet processing system DPDK [4]), and DPDK vSwitch [5], a port of OvS that uses DPDK. We focus our measurements on OvS because it is the latest and fastest forwarding method based on the Linux network stack and the existence of the DPDK port allows for a direct performance comparison of the Linux network stack with DPDK. Thus we can show where potentially unnecessary bottlenecks in kernel-based packet processing systems are.

The throughput of DPDK-based software is significantly faster than the kernel forwarding techniques. We use profiling techniques to understand why the kernel applications are slower. We analyze hardware bottlenecks like effects of the CPU cache and software bottlenecks in the applications and the kernel. Based on these results we conclude that the most important bottleneck is receiving and sending packets in the network stack and that a six-fold performance improvement for OvS can be achieved by replacing the I/O technique with DPDK or a similar framework like Netmap [6] or PF_RING DNA [7].

We begin with a description of our test setup in Section II. Section III presents the results of our throughput tests. Section IV discusses potential hardware bottlenecks and Section V software bottlenecks. We discuss related work in Section VI and conclude with an outlook.

II. Test Methodology

Our test setup in Fig. 1 is based on recommendations by RFC 2544 [8].

A. Hardware Setup

Servers A and B are used as load generators and packet counters, the DuT (Device under Test) runs the software under test. For black-box tests, we must not introduce any overhead on the DuT through measurements. So we measure the offered load and the throughput on A and B. The DuT runs the Linux tool `perf` for white-box tests; this overhead reduces the maximum throughput by $\sim 1\%$.

The DuT uses an Intel X520-SR2 dual 10 GbE network interface card (NIC), the two other servers are equipped with X520-SR1 single port NICs. These NICs are based on the Intel 82599 Ethernet controller. All servers use 3.3 GHz Intel Xeon E3-1230 V2 CPUs. We also tested a second setup in which we replaced the X520 NICs with newer Intel X540 NICs to test effects of different hardware. We disabled Hyper-Threading, Turbo Boost, and power saving features that scale the frequency with the CPU load because we observed measurement artifacts with these features.

B. Software Setup

As generation of 64 B packets at 10 GbE line rate is a task that many existing packet generators are incapable of, we used a modified version of the `pfsend` packet generator from the PF_RING DNA [7] software repository [9]. This packet generator is able to produce UDP packets at the 10 GbE line rate of 14.88 Mpps with only a single core. Except for the DPDK forwarding test, all tests just used unidirectional traffic because the line rate was not the bottleneck.

We restrict the tests to a single flow and CPU core, because we observed linear scaling with the number of available CPU

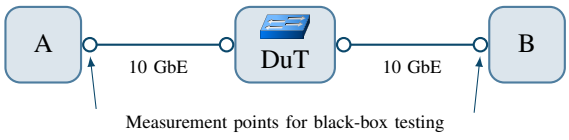


Figure 1. Server setup

 ISBN: 978-1-61208-398-8

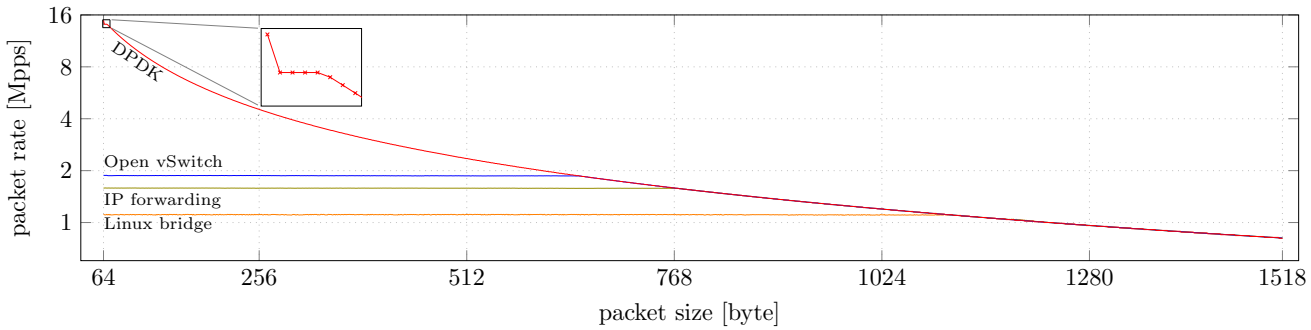


Figure 2. Packet size vs. throughput (logarithmic y-axis)

cores in previous work [10]. This focus on packet processing per core allows us to make claims for systems with different numbers of cores.

We used our own packet counter that relies on the statistics registers of our NICs which are accessible via `ethtool`. The packet rate is calculated by snapshotting the NIC counters periodically.

The DuT runs the Debian-based live Linux distribution Grml with a 3.7 kernel, the `ixgbe` 3.14.5 NIC driver with interrupts statically assigned to CPU cores, OvS 2.0.0, DPDK vSwitch 0.10 (based on OvS 2.0.0), and Intel DPDK 1.6.0.

C. Presentation of Results

All throughput measurements were run for 30 seconds and the packet rate was sampled every 100 ms. Graphs show the average measurement. The standard deviation was below 0.2% for all throughput measurements. We therefore omitted error bars in these. Profiling measurements were restricted to the core on which the processing task was pinned and were run for five minutes per test to get accurate results. Measurements showing significant noise were plotted with 95% confidence intervals (cf. Fig. 4).

III. Forwarding Performance

Table I compares the data plane performance of OvS, Linux IP forwarding, Linux bridge, DPDK vSwitch, and DPDK L2FWD with minimally sized packets. The packet size is irrelevant in almost all scenarios as shown in Fig. 2.

TABLE I. Data plane performance comparison

Application	Throughput [Mpps]
DPDK L2FWD bidir X540	29.76
DPDK L2FWD bidir X520	24.06
DPDK L2FWD unidir X520	14.88
DPDK vSwitch	11.31
Open vSwitch	1.88
Linux IP forwarding	1.58
Linux bridge	1.11

The OvS kernel module is able to process packets faster than the Linux kernel forwarding. The Linux kernel code for routing has received steady optimizations while the bridging code was last modified with kernel 2.6. OvS proved to be the fastest packet forwarding technique using the Linux network stack.

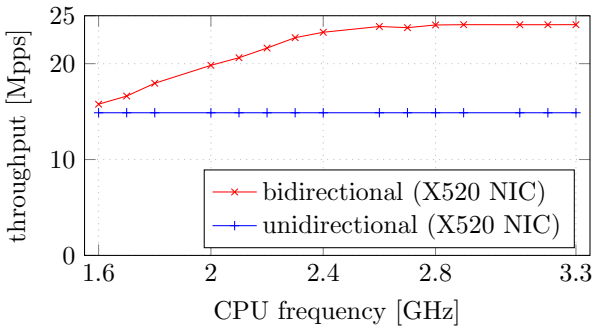


Figure 3. L2FWD at different clock rates

The DPDK applications do not use the Linux network stack and are significantly faster. The DPDK port of OvS showed a six-fold performance increase compared to the kernel version.

A. User Space Packet Processing

Approaches like Intel DPDK [4], Netmap [6], and PF_RING DNA [7] replace the network stack with a user space application to avoid overhead.

DPDK L2FWD only forwards packets between two statically configured network interfaces without consulting a routing or flow table. It can therefore be seen as an upper bound for the possible throughput. We focus our measurements on DPDK here but also observed similar results with forwarding applications based on Netmap and PF_RING.

We adjusted the CPU clock frequency to measure the required processing power per packet. DPDK L2FWD managed to forward 14.88 Mpps even with the lowest possible frequency, we therefore added a bidirectional test for this application. Fig. 3 shows the throughput with all clock frequencies supported by our CPU. The same test with similar results for the performance of Netmap is presented by Rizzo in [6].

The bidirectional test initially only achieved a throughput of 24.06 Mpps instead of line rate with the maximum clock frequency. We then tried to use two cores for this test but this resulted in the same performance which indicates a hardware limit in the NIC. We therefore replaced the X520 NIC with a newer X540 NIC on all servers to investigate further. The X540 was able to forward 29.76 Mpps, i.e., line rate, with DPDK on a single CPU core.

The DPDK L2FWD application initially only managed to forward 13.8 Mpps in the single direction test at the maximum

 ISBN: 978-1-61208-398-8

CPU frequency, a similar result can be found in [11]. Reducing the CPU frequency increased the throughput to the expected value of 14.88 Mpps. Our investigation of this anomaly revealed that the lack of any processing combined with the fast CPU caused DPDK to poll the NIC too often. DPDK does not use interrupts, it utilizes a busy wait loop that polls the NIC until at least one packet is returned. This resulted in a high poll rate which affected the throughput. We limited the poll rate to 500,000 poll operations per second (i.e., a batch size of about 30 packets) and achieved line rate in the unidirectional test with all frequencies. This effect was only observed with the X520 NIC, tests with X540 NICs did not show this anomaly.

IV. Hardware Bottlenecks

We use OvS as an example and follow a packet's path through it and examine each component for potential bottlenecks. A packet arrives at the input network interface and is transferred via DMA with Intel's Direct Cache Access (DCA) technology [12] to the L3 cache. It is then processed and modified by OvS on the CPU based on a flow table in the OvS kernel module called the *datapath*. Packets that do not match any rule in the datapath are forwarded to a user space process, which then installs a rule in the kernel module for subsequent packets of this flow, this processing path is called the slow path. These rules use an idle timeout so that only actively used rules are kept in the kernel module. Afterwards the packet is transferred to the outgoing network interface via DMA/DCA.

Other forwarding systems beside OvS use the same packet flow except for the processing step. The following potential bottlenecks are present in the packet processing path.

A. Network Bandwidth

DPDK L2FWD hit the limit of 14.88 Mpps in the unidirectional test, but not in the bidirectional test (cf. Table I). All other measured programs were far below this limit. It is therefore not a relevant bottleneck for our tests with a single CPU core.

B. NIC Processing Capacity

The data sheet of the Intel 82559 chip does not mention any limits to the packet rate [13]. However, we have encountered such a limit at 24 Mpps. We verified that the processing is limited by the number of packets per second and not the total bandwidth by testing with larger packet sizes. The NIC is able to handle line rate with packets larger than 100 Byte. The newer X540 chip does not have this limit.

C. PCIe Bandwidth

Both the X520 and X540 NICs we used are attached via a PCIe 2.0 x8 link with a net bandwidth of 32 GBit/s per direction, far more than the 20 GBit/s of the two network ports. Our CPU unfortunately does not support performance counters to measure this bandwidth. However, this limit is not relevant since the X540 NIC is able to sustain full line rate with 64 Byte packets on both ports.

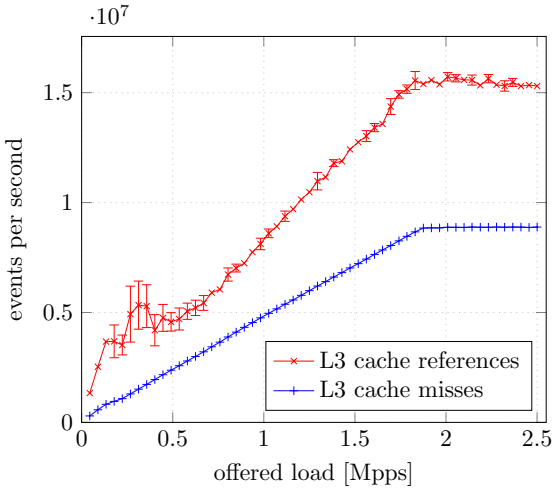


Figure 4. L3 cache statistics (OvS)

D. Memory Bandwidth

Each packet is written to and read from the main memory at least once. The CPU in our test server offers a memory bandwidth of 200 GBit/s. This is therefore not a bottleneck for 10 GBit networks but needs to be considered when moving to 100 GBit Ethernet.

E. CPU Cache Size

The overall cache size can be a bottleneck if it is insufficient for the state of the application.

We use OvS as example here, but the results are also applicable to other forwarding applications, which use a flow table or a routing table. Fig. 4 shows the number of L3 cache references and misses in the OvS forwarding scenario with only one flow. Both grow linearly with the number of processed packets per second, the miss ratio stays constant. Slight deviations in the lower packet rates are due to the dynamic interrupt rate throttling by the ixgbe driver. The other cache levels show similar results.

The total number of accesses and misses per second is in the order of 10^7. This translates to a cache and memory bandwidth of less than 8 GBit per second when multiplied with the CPU cache line size of 64 Bytes. This is an uncritically low bandwidth requirement that can easily be satisfied [14].

A high number of actively used flow table entries, which are 576 Bytes each, in the OvS forwarding scenario can exhaust the cache. The L1 cache fits 56 entries, the L2 cache 455, and the L3 Cache about 14 500 without taking space requirements for packets or any other required data into account.

Fig. 5 shows the number of active flows vs. cache misses and throughput. The first two caches quickly fill up and cause a slight drop in the performance from 1.87 Mpps with one flow (slightly lower than the result from Table I due to active profiling) to 1.76 Mpps with 2000 flows.

Tests that exhaust the L3 cache require more than 14 500 flows, but testing with such a large amount of flow table entries was not feasible due to exponential growth of the time required to add flows. This exponential growth of flow table

 ISBN: 978-1-61208-398-8

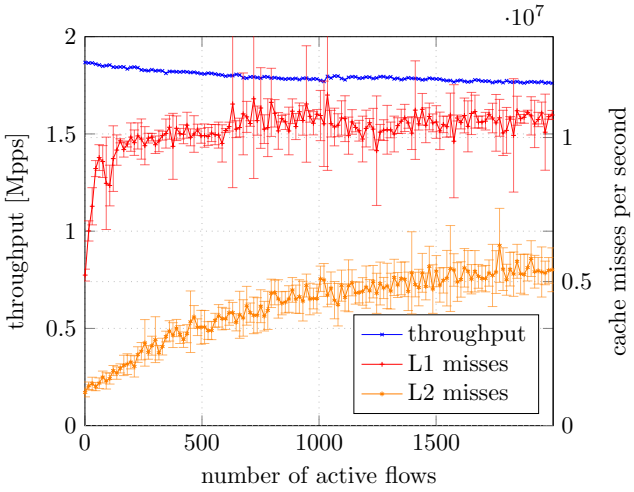


Figure 5. Flow table size vs. cache misses (OvS)

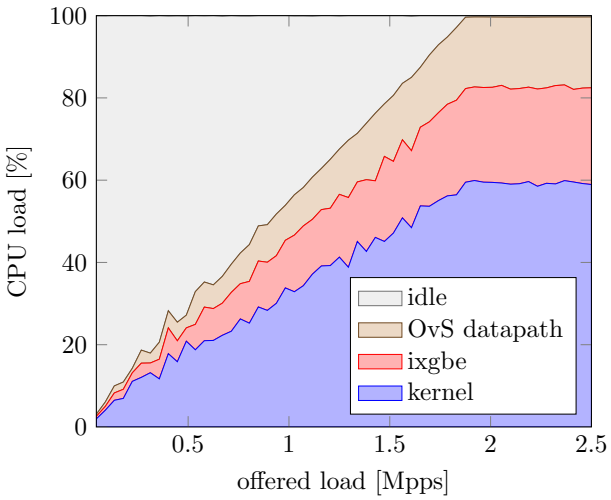


Figure 6. CPU usage per kernel module (OvS)

modification operations in OvS is described in more detail by Rotsos et al. [15]. Note that this is not a bottleneck in a real-world application because OvS only needs a constant amount of flow table entries (exact figure depends on the installed OpenFlow rules) per attached device due to wild card support.

Minor performance improvements can be achieved by reducing the footprint of flow or routing table entries. But this is not a major bottleneck and does not explain the performance gap to DPDK-based applications.

F. CPU Cache Line Length

The cache line length can also affect the throughput due to access latencies and bad alignment when using packet sizes that are not a multiple of the cache line length of 64 Bytes [14]. To investigate we performed a test series of maximum throughput experiments and we increased the packet size by 1 Byte per test (cf. Fig. 2).

The DPDK L2FWD throughput test showed a very slight deviation from the line rate for packet sizes between 65 and 68 Bytes, which were processed with only 14.12 Mpps instead of the expected 14.71 to 14.20 Mpps (line rate). Larger packets are limited by the line rate. We assume this is caused by packets that need slightly more than a single cache line. Rizzo measured this effect with Netmap in more detail in [6].

The maximum throughput curves of the conventional packet processing systems have no inflection points except when the network link bandwidth sets in. We conclude that there are no adverse effects if the packet sizes do not match the cache line size for the kernel-based systems which we are trying to improve.

G. CPU Time

We measured the throughput of all packet forwarding programs with different CPU frequencies to analyze the impact of raw processing power. All applications scaled linearly with the CPU frequency except for the DPDK L2FWD application (cf. Section III-A). This means that the only relevant bottleneck for the kernel-based forwarding applications is the software.

V. Software Bottlenecks

We ran the Linux profiling tool `perf` to analyze the relative time required per function and combined this with the CPU cycle counter to compute the wall-clock time of each function.

A. CPU Utilization per Processing Step

Fig. 6 shows the CPU time aggregated per kernel module in the OvS forwarding scenario under increasing load. It shows the self-time spent in the respective kernel module, i.e., without taking the call stack into account. The load of all involved modules increases linearly and the first drops were observed once the CPU load hit 100%. This example uses OvS because its architecture as a separate kernel module allows for a simple split by functionality, the other kernel-based forwarding applications show a similar behavior. DPDK utilizes a busy wait loop and the CPU load is always 100% independent of the offered load.

The OvS datapath module is responsible for the forwarding decision, the other modules handle all other tasks: receiving packets, sending packets, and buffer management. Only about 18% of the CPU time is spent in the OvS kernel module, so the software bottleneck is packet I/O and not processing.

The datapath module calls back into other kernel modules during the processing at one point to acquire and release a spin lock which should be attributed to OvS even though the time is spent in another kernel module. Measuring this requires a kernel that is compiled without the `fomit-frame-pointers` compiler optimization to generate backtraces based on the stack from a sample. This change resulted in a drop of the throughput by $\sim 15\%$. This additional overhead is distributed approximately uniformly across the code which we confirmed by comparing the self-times with and without the compiler optimization. The option allows for a more detailed and correct look at the CPU time distribution: about 4% of the total time is spent in the spin lock with OvS as caller, so 22% of the time is spent processing the packet and 78% are for receiving and sending. All of the following measurements were run without this compiler optimization.

 ISBN: 978-1-61208-398-8

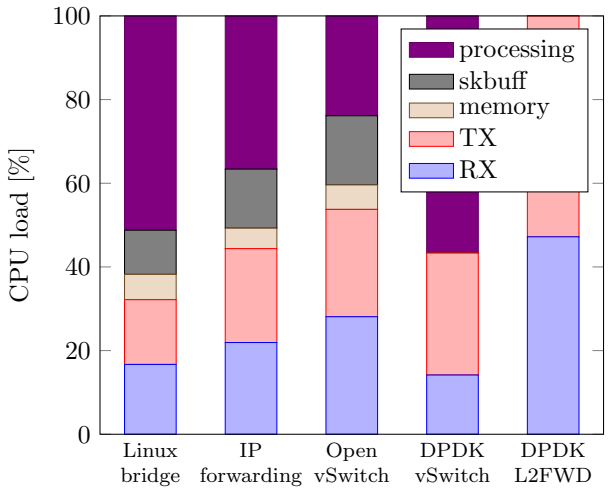


Figure 7. CPU usage per task under full load

Fig. 7 shows the relative required time for the required steps of the forwarding methods: `Processing` are all functions that are associated with reaching a forwarding decision, so it is 0% for the DPDK L2FWD application which only transfers packets between two interfaces. `Memory` represents the time required to allocate and free the memory for the packets, but without any initialization functions that are specific to the kernel's skbuff data structure. `Skbuff` contains all functions that initialize, release, or change an skbuff (kernel packet buffer descriptor) without the memory management functions. `RX` includes all functions that handle receiving a packet in the network stack and driver. It also includes time required to handle interrupts like saving and restoring the context or raising the software interrupt to call into `softirqd` for further processing. `TX` includes all functions from the point where the application calls into the network stack's `dev_queue_xmit` transmit function. It includes any task in the network stack that is required to send a packet, most notably the handling of the queuing discipline in the kernel.

Table II shows the same data as Fig. 7 in CPU cycles per packet. The Linux bridge clones the skbuff descriptor unnecessarily which leads to higher memory management and skbuff times. The differences between the I/O tasks for the DPDK applications are due to different batch sizes which were chosen for optimal overall performance with the applications.

TABLE II. CPU CYCLES PER PACKET AND PROCESSING STEP

Application	RX	TX	Mem.	skb	Process.
Linux bridge	491	456	178	311	1508
IP forward	453	463	101	292	757
Open vSwitch	490	447	102	288	416
DPDK vSwitch	41	85	0	0	165
DPDK L2FWD	55	62	0	0	0

The linux kernel needs about 1300 cycles on average to receive and send a single packet including memory management and overhead from the data structures. DPDK performs the same tasks with only 120 cycles, so the network stack is an important bottleneck for packet forwarding systems.

B. Overheads in the Kernel

The Linux network stack is a general-purpose network stack and exhibits unnecessary overheads when used with a specialized application like a packet forwarding system. One example for such an overhead in the kernel is the send path which requires a spin lock to access the queuing discipline of the device (configured as the default fifo queue). This lock takes about 8% of the total wall-clock time in the OvS example which impacts the throughput. Profiling shows that this lock is never contended because there are no other network tasks which could acquire the same lock. However, the kernel needs to handle the generic case which might be configured differently, so the spin lock still needs to be acquired and released at multiple locations, this manifests as overhead.

Another example is the skbuff data structure. It needs to handle all possible cases for the network stack and requires an extensive constructor and destructor. A simple buffer is sufficient for a software switch or router.

A switch can allocate a fixed amount of buffers on startup and directly reuse them after finishing a batch of packets. The kernel's network stack uses a memory pool from which buffers are acquired before retrieving packets and returned to after sending packets. In the general case, this is necessary because packets might still be needed after processing them. However, a software switch can always forward all packets directly. So the same buffers can be overwritten with new packets immediately after sending a batch. Memory management is therefore also unnecessary.

C. Open vSwitch vs. DPDK vSwitch

DPDK vSwitch shows a six-fold increase in throughput over OvS, this discrepancy can not explained by only the packet I/O. The processing logic is also optimized, DPDK vSwitch only takes 165 cycles to reach a forwarding decision for a packet while the original OvS requires 416 cycles.

DPDK vSwitch replaces the whole flow table with a highly optimized version. Some of these optimizations, like an optimized hash function that utilizes the CPU's CRC32 instruction, could be ported back to the original OvS. This would improve the performance slightly, but the network stack is still the main bottleneck.

VI. Related Work

Bolla and Bruschi recognized the trend for optimization already in 2007 and presented a detailed study of packet forwarding in Linux by applying RFC 2544 and internal measurements with profiling [16]. They measured a higher CPU load of packet I/O tasks in test similar to the one described in Section V here. This indicates that the overhead was reduced since kernel version 2.6.16 which they used. Later a study of performance influences of multi-core PC systems under different workloads [17] was published. These performance studies have been shown and extended, e.g., in [10]. We only discuss bottlenecks of software routers here, further measurements on Open vSwitch throughput and latency in different scenarios can be found in [18] and [19].

Another important aspect beside the throughput is the latency of a forwarding system. A notable example of latency

 ISBN: 978-1-61208-398-8

measurements is the OFLOPS OpenFlow benchmarking framework by Rotsos et al. which uses OvS as an example, they also describe challenges with measuring latency on commodity hardware [15]. Discussion of latency in software routers can also be found in [20] and [21]. We will extend this study to include latency measurements based on our packet generator MoonGen [22] in future work.

Other user space packet processing frameworks beside Intel's DPDK with similar performance characteristics are Netmap by Rizzo who presents similar measurements to our tests from Section III-A for his framework [6]. This shows that DPDK and Netmap have similar performance characteristics and our results are transferable to Netmap. Deri presents PF_RING DNA in [7] which was originally written for a fast packet capture application called nCap. We expect similar performance gains when using this framework.

DPDK was also used to evaluate the performance of a PC-based OpenFlow switch by Pongracz et al. [11] but without comparing it to a conventional packet I/O system like the Linux kernel.

Rizzo et al. ported the networking library libpcap to Netmap and use it to improve applications transparently with a user space version of OvS as one example [23]. They only state the performance improvements for various applications but do not measure software bottlenecks explicitly.

VII. Conclusions and Outlook

We identified and measured different bottlenecks and conclude that the main bottleneck for PC-based packet forwarding systems is the software due to overhead in the kernel's network stack. We measured that receiving and sending packets with a user space packet processing framework like DPDK is 12 times faster than using the Linux kernel to do the same task. Existing software switches like OvS can show significant performance improvements by adopting a modern packet I/O framework even without modifying the processing logic. DPDK vSwitch also optimizes the processing logic and shows a six-fold increase of the total performance compared to OvS.

There have been discussions to include the Netmap framework in the Linux kernel [24] where it could supplement the current network API while maintaining backwards compatibility with older drivers and applications. Using such a framework requires special drivers and a complicated setup procedure at the moment. Direct support from the Linux kernel is an important step for the mainstream adaption of such frameworks in the next generation of software switches and routers.

ACKNOWLEDGMENTS

This research has been supported by the DFG as part of the MEMPHIS project (CA 595/5-2), the KIC EIT ICT Labs on SDN, and the BMBF under EUREKA-Project SASER (01BP12300A).

References

[1] L. Rizzo and G. Lettieri, "VALE, a Switched Ethernet for Virtual Machines," in Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies. ACM, 2012, pp. 61–72.

[2] "Vyatta," http://www.brocade.com/products/all/network-functions-virtualization/product-details/5400-vrouter/index.page, last visited 2015-03-08.

[3] "Open vSwitch," http://openvswitch.org, last visited 2015-03-08.

[4] "Intel DPDK: Data Plane Development Kit," http://dpdk.org/, Intel, last visited 2015-03-08.

[5] "Intel DPDK vSwitch," https://github.com/01org/dpdk-ovs, Intel Corporation, last visited 2015-03-08.

[6] L. Rizzo, "Netmap: A Novel Framework for Fast Packet I/O," in 2012 USENIX Annual Technical Conference (USENIX ATC 12). Boston, MA: USENIX, 2012, pp. 101–112.

[7] L. Deri, "nCap: Wire-speed Packet Capture and Transmission," in IEEE Workshop on End-to-End Monitoring Techniques and Services, 2005, pp. 47–55.

[8] S. Bradner and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices," RFC 2544 (Informational), Internet Engineering Task Force, March 1999.

[9] "Ntop," http://www.ntop.org, last visited 2015-03-08.

[10] T. Meyer, F. Wohlfart, D. Raumer, B. Wolfinger, and G. Carle, "Validated Model-Based Prediction of Multi-Core Software Router Performance," Praxis der Informationsverarbeitung und Kommunikation (PIK), April 2014.

[11] G. Pongracz, L. Molnar, and Z. L. Kis, "Removing Roadblocks from SDN: OpenFlow Software Switch Performance on Intel DPDK," Second European Workshop on Software Defined Networks (EWSDN'13), 2013, pp. 62–67.

[12] R. Huggahalli, R. Iyer, and S. Tetrick, "Direct Cache Access for High Bandwidth Network I/O," in Proceedings of the 32nd Annual International Symposium on Computer Architecture, 2005, pp. 50–59.

[13] "Intel 82599 10 GbE Controller Datasheet Rev. 2.76," Intel, 2012, Santa Clara, USA.

[14] "Intel® 64 and IA-32 Architectures Optimization Reference Manual," Intel, 2014.

[15] C. Rotsos, N. Sarrar, S. Uhlig, R. Sherwood, and A. W. Moore, "Oflops: An Open Framework for OpenFlow Switch Evaluation," in Passive and Active Measurement. Springer, 2012, pp. 85–95.

[16] R. Bolla and R. Bruschi, "Linux Software Router: Data Plane Optimization and Performance Evaluation," Journal of Networks, vol. 2, no. 3, June 2007, pp. 6–17.

[17] M. Dobrescu, K. Argyraki, and S. Ratnasamy, "Toward Predictable Performance in Software Packet-Processing Platforms," in USENIX Conference on Networked Systems Design and Implementation (NSDI), April 2012.

[18] A. Beifuß, D. Raumer, P. Emmerich, T. M. Runge, F. Wohlfart, B. E. Wolfinger, and G. Carle, "A Study of Networking Software Induced Latency," in 2nd International Conference on Networked Systems 2015 (accepted), Cottbus, Germany, 2015.

[19] P. Emmerich, D. Raumer, F. Wohlfart, and G. Carle, "Performance Characteristics of Virtual Switching," in 2014 IEEE 3rd International Conference on Cloud Networking (CloudNet), Luxembourg, 2014.

[20] L. Angrisani, G. Ventre, L. Peluso, and A. Tedesco, "Measurement of Processing and Queuing Delays Introduced by an Open-Source Router in a Single-Hop Network," IEEE Transactions on Instrumentation and Measurement, vol. 55, no. 4, 2006, pp. 1065–1076.

[21] S. Larsen, P. Sarangam, R. Huggahalli, and S. Kulkarni, "Architectural Breakdown of End-to-End Latency in a TCP/IP Network," International Journal of Parallel Programming, vol. 37, no. 6, 2009, pp. 556–571.

[22] P. Emmerich, F. Wohlfart, D. Raumer, and G. Carle, "MoonGen: A Scriptable High-Speed Packet Generator," ArXiv e-prints, Oct. 2014. [Online]. Available: http://adsabs.harvard.edu/abs/2014arXiv1410.3322E

[23] L. Rizzo, M. Carbone, and G. Catalli, "Transparent Acceleration of Software Packet Forwarding using Netmap," in INFOCOM, 2012 Proceedings IEEE. IEEE, 2012, pp. 2471–2479.

[24] S. Hemminger, "netmap: infrastructure (in staging)," http://lwn.net/Articles/548077/, last visited 2015-03-08.

 ISBN: 978-1-61208-398-8

Developing a Simulator Applied to Audio Coding Process MPEG-4 AAC

Mauricio Harff, Marcio Garcia Martins, Arthur Tórgo Gómez
Postgraduate Interdisciplinary Program in Applied Computing
University of Vale do Rio dos Sinos
São Leopoldo, Brazil
e-mail: mharff@gmail.com, marciog@unisinos.br, breno@unisinos.br

***Abstract*—This paper proposes the development of a simulator to find the internal parameters of the MPEG-4 Advanced Audio Coding (AAC), so as to optimize the perceptual audio quality for a given bit rate. The implementation of the simulator was developed in ANSI C programming language, using the Tabu Search and Genetic Algorithm in a hybrid structure. Through the minimization of the Average Noise-to-Mask Ratio (ANMR) metric, the simulator identifies the best configuration of internal parameters of the MPEG-4 AAC and improves the perceptual audio quality.**

Keywords-audio compression; MPEG-4 Advanced Audio Coding; Metaheuristics.

I. INTRODUCTION

During the encoding process, the MPEG-4 AAC encoder must dynamically choose the parameters that determine the best perceptual audio quality. This process of determining the best parameters was defined as the AAC Encoding Problem [1]. In this process, the reference encoder uses an iterative method known as Two Loop Search (TLS) [2] to define the parameters for a particular audio frame. However, this search method of parameters used, with respect to encoder MPEG-4 AAC, does not solve optimally the AAC Encoding Problem [3]. This paper proposes a simulator applied to the AAC encoder, based on a hybrid algorithm of metaheuristics Tabu Search (TS) [4] and Genetic Algorithm (GA) [5], that through experiments may define the parameters through the obtaining of solutions of good quality. Thus, the simulator proposed realizes the function of the dynamic definition of internal parameters of encoder configuration.

This paper is organized as follows. Section 2 presents related work. Section 3 presents and discusses the MPEG-4 encoder. Section 4 presents the mathematical formulation proposed. Section 5 shows the computational architecture of the simulator. Section 6 presents the results obtained. And finally, Section 7 presents the conclusions.

II. RELATED WORK

In the literature, one can find some works that address the AAC Encoding Problem, using different solution techniques and some variations in the structure of the solution model. These approaches may simplifying the model, proposing simpler methods to solve it, or make the model more faithful to reality, by the addition of decision variables and generate results with higher perceptual audio quality. The problem with the addition of more decision variables is that the problem becomes more complex to solve, and so mathematical methods of solution require a higher computational time for its resolution. In the work of Aggarwall [2], an algorithm called Search Trellis, based on a trellis arrangement is introduced to improve the efficiency of the AAC encoder. This paper proposes a quantization step optimization in the encoder, so that get produce at lower bit rates when compared to the reference encoder. In [2], were used only two decision variables, the Scale Factors (SF) and Huffman Code Books (HCB). The model takes into account in the Objective Function (OF), the amount of bits needed to represent the quantized spectral coefficients and also the side information while preserving the perceptual quality of the audio, i.e., minimizing the ANMR metric. The results obtained in the Aggarwall's work show an improvement in perceptual audio quality, compared to the reference encoder to a same bit rate. The improvement achieved resulted in an increase of two times in the perceptual audio quality, which represents a decrease twice in the ANMR metric.

Continuing the work of Aggarwall [2], an improvement was proposed by Bauer [3]. Even using only two decision variables, the SF and HCB, the author proposed a different technique to solve the AAC Encoding Problem. The method, called Fast Trellis Search, provides a perceptual quality very close to Trellis Search algorithm, but solves the AAC Encoding Problem approximately 25 times faster. Due to the computational time needed to solve the problem, the old technique was unable to make an implementation of AAC encoder in real time.

In [2] and [6], the authors, also discuss the AAC Encoding Problem. However, these works use not only two decision variables in the model, and so the problem becomes more true to life and at the same time more complex. Using Mixed Integer Linear Programming (MILP), the problem is modeled by making use of four decision variables: blocks structure, grouping of blocks, SF and HCB. The mathematical model that describes the problem to four decision variables is complex and extensive. In this approach, the authors are able to obtain significant improvements in perceptual audio quality when compared with the structures of the reference encoding models and the others discussed before. However, due to the complexity of the problem, the computational processing time becomes very high. This makes it impossible an encoder with such structure of parameter optimization to be implemented in real time.

 ISBN: 978-1-61208-398-8

III. MPEG-4 AAC ENCODER

The dynamics of the MPEG-4 AAC encoder starts with data processing, the conversion of the audio signal in time domain to the frequency domain using the Modified Discrete Cosine Transform (MDCT). After this procedure, the coefficients of a signal frame are divided into blocks, may be formed one long block or 8 short blocks [7]. If the frame coefficients are divided into 8 short blocks, they can still be combined into sections to reduce the total bit rate. Within a long block or sections, spectral coefficients are divided again, however in the frequency domain, in Scale Factor Band (SFB). For each SFB, the encoder sets an amount of bits to represent the frequency coefficients. This procedure occurs dynamically, in real time, through the iteration of bit rate control loop and the distortion control loop [2]. These loops define the quantization interval and the HCB that better configure the encoder to a SFB of a frame. Furthermore, it is necessary that the encoder transmits in its output data flow, the value of SF and HCB associated to each SFB, as side information [8]. The encoder also uses some bits, to tell the decoder which the grouping used in the sections of the blocks. In this way, the overall bit rate needed to encode, a given audio signal, is the sum of the bits needed to the coding of the sections and of the blocks, with the sum of the bits needed to encode side information. Therefore, the four parameters to be managed by the MPEG-4 AAC encoder can be defined [3]:

- Blocks Structure (short block, long block);
- Blocks Grouping (only for short blocks);
- Quantization Interval (Scale Factors);
- Entropy Encoder (Huffman Code Books).

To ensure a low bit rate, the encoder must choose a good combination of the four parameters, so that the predefined bit rate not is exceeded and that the perceptual audio signal quality is guaranteed. To measure the perceptual encoded signal quality, the ANMR metric is used. This is a metric that incorporates the human ear model in its conception [9]. The noise or quantization errors for specific frequency bands (critical bands) [9] are computed, so that the level of importance of noise at a given frequency range is obtained through the response curve of the human ear model [10]. In this way, it is possible to represent the perceptual audio quality through an exact parameter [11].

In [1][3] and [6], the authors define the ANMR optimization process as the AAC Encoding Problem. This problem resembles a classical combinatorial optimization problem, the Part Selection Problem. The problem of selecting parts belongs to the group of NP-Complete complexity problems [10].

Figure 1 shows the architecture of the MPEG-4 AAC encoder, with the simulator structure based on metaheuristics inserted in the diagram.

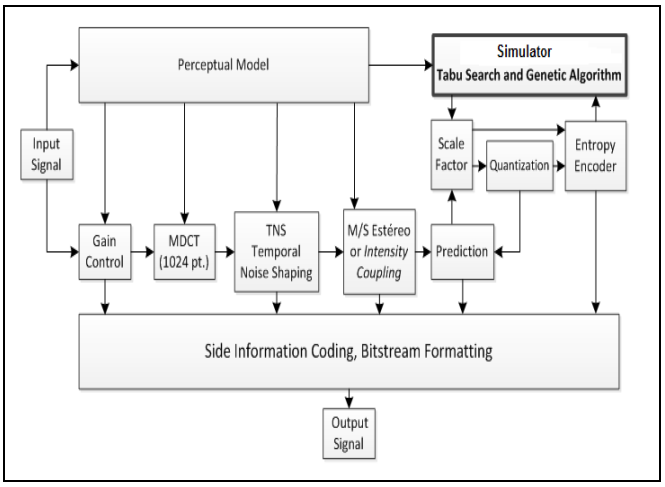


Figure 1. The architecture of the MPEG-4 AAC encoder with the proposed simulator.

Find the optimal solution to the problem of AAC encoding is a complex optimization problem, due to the interdependencies of the four parameters: blocks structure, blocks grouping, and quantization interval and entropy encoder [3]. The approaches found in the literature about the resolution of the AAC Encoding Problem, do not allow real time implementations. Thus, it is necessary simplify the problem. The AAC Encoding Problem is significantly simplified if the relationship between some parameters is neglected, as for example, the interdependence of adjacent Scale Factors, which are encoded differentially.

The method proposed by the standards [7][8] to solve the AAC Encoding Problem, is the iterative method TLS. However, some important disadvantages of the TLS procedure are that it does not necessarily converge, ignores the inter-band correlations of Scale Factors and Huffman Code Books, and is not intended to minimize the total distortion due to consider their analysis separately by bands [11] [12].

IV. MATHEMATICAL FORMULATION

The mathematical model, used in this paper, is based on the approach of Claus Bauer [3]. However, are considered three parameters of the problem: the quantization intervals, represented by SF, the HCB for each SFB, and the grouping of blocks for each frame, if it is configured as short block. The choosing process of the type of block, between short block and long block, is responsibility of the MPEG-4 AAC, because it uses the Perceptual Entropy metric [10], which identifies the minimum number of bits required to represent the frame. Thus, the objective of the simulator is to find the best combination of the three parameters, for a given bit rate pre-defined, so that the perceptual audio quality is maintained or improved.

Due to the mathematical formulation of the AAC Encoding Problem be very extensive, with several expressions to represent the constraints, in this paper, the mathematical model will be represented in a summary form. In [3], the full model is presented.

 ISBN: 978-1-61208-398-8

$Minimize$

$$ANMR(Z) = \frac{1}{N}\sum_{g=0}^{G+1}\sum_{i=1}^{8R}\sum_{a=1}^{M_1}\sum_{b=1}^{M_2} e_{g,i,a} Z_{g,i,a,b} \quad (1)$$

$Subject\ to$:

$$R(Z, A, B, g, U, X) \le R_t\ , \quad (2)$$

where

$$R(Z, A, B, g, U, X) = \quad (3)$$

$$= \sum_{g=0}^{G+1}\sum_{i=1}^{120}\sum_{a=1}^{M_1}\sum_{b=1}^{M_2} Z_{g,i,a,b} Q_{g,i}(a,b) \quad +$$

$$+ \sum_{i=2}^{T}\sum_{a=0}^{2M_1} A_{L,i,a} F^*(a) + \sum_{i=2}^{120}\sum_{a=0}^{2M_1} A_{S,i,a} F^*(a) +$$

$$+ \sum_{i=2}^{T}\sum_{b=0}^{2M_2} B_{L,i,b} G^*(b) + \sum_{i=2}^{120}\sum_{b=0}^{2M_2} B_{S,i,b} G^*(b) \quad +$$

$$+ 4x_1 + 3\sum_{j=0}^{7}(x_{2,j} + x_{3,j} + x_{4,j}) + c\sum_{j=1}^{7} u_j\ ,$$

where:

- N = Total number of Scale Factor Bands within a frame;
- R = Total bit rate;
- R_t = Total maximum bit rate specified;
- i = The i-th Scale Factor Band;
- j = The j-th part of a frame encoded as short block;
- g, G = Type of Block and Block Grouping;
- a, A = Scale Factor (SF);
- b, B = Huffman Code Books (HCB);
- $Z_{g,i,a,b}$ = Specific parameters of quantization error;
- $e_{g,i,a}$ = Distortion weight for a specific configuration;
- $Q_{g,i}(a,b)$ = Number of bits required to encode spectral coefficients for a particular configuration of SF and HCB;
- M_1 = Highest value admissible for the SF;
- M_2 = Highest value admissible for the HCB;
- L = Parameters associated with a long block;
- S = Parameters associated with a short block;
- $F^*() =$ Function that represents the number of bits to differentially encode the SF: $a_i - a_{i-1}$ (side information);
- $G^*() =$ Function that represents the number of bits for encoding the HCB: $b_i - b_{i-1}$ (side information);
- x, X = Auxiliary variables for calculating the quantity of bits (side information);
- U, u = Number of bits to represent the grouping of blocks (side information).

Equation (1) is the Objective Function (OF) of the AAC Encoding Problem that must be minimized. This function is a weighted sum of quantization errors (distortion) of audio signal in each Scale Factor Band (SFB). The quantization error is defined by $Z_{g,i,a,b}$ for the i-th SFB, to a configuration g of blocks, where the i-th SF is chosen equal to a and the i-th HCB equal to b. The inverse of the masking threshold, which defines the behavior of the human ear, is defined by the weight $e_{g,i,a}$ of the i-th SFB in the g-th configuration. In this way, the lower the value of the sum of ANMR, the greater the perceptual audio signal quality.

The transmission rate is the main constraint on the model of AAC encoder, because once determined by the user, this bit rate should not be exceeded. The maximum bit rate defined by the user is represented by R_t in (2). Thus, the bit rate produced by the encoder should always be less than the rate R_t.

The constraint (3) represents every portion of the bit rate, which together, comprise the total bit rate stream output of MPEG-4 AAC encoder. The first term of summations in (3) represents the amount of bits used to encode the spectral coefficients. The following two summations in (3) represent the bit rate to encode the Scale Factors of long blocks and short blocks respectively. Analogously, the following two summations represent the amount of bits to encode the Huffman Code Books, used in long and short blocks respectively. The last two summations represent the number of bits used to encode the arrangement structure of blocks, as side information. This side information is fundamental since it allows the decoder interpret how the data are organized within the data stream.

V. SIMULATOR ARCHITECTURE

The simulator architecture is composed by three modules: Initial Solution, Genetic Algorithm and Tabu Search. Thus, at the end of the implementation of metaheuristics, the best solution obtained to the configuration of the internal parameters is delivered to the MPEG-4 AAC encoder. It should be noted that this structure is performed individually for each audio frame.

Figure 2 shows the hybrid architecture developed to solve the AAC Encoding Problem, which is explained below.

Step 1: The initial solution used by the Genetic Algorithm is generated during the TLS iteration existing within the MPEG-4 AAC reference encoder. So, after run an iteration of external loop and an iteration of internal loop of the TLS algorithm [8], a solution is obtained. The solution still is randomly modified, with probability 0.05, before inserted into the initial population.

 ISBN: 978-1-61208-398-8

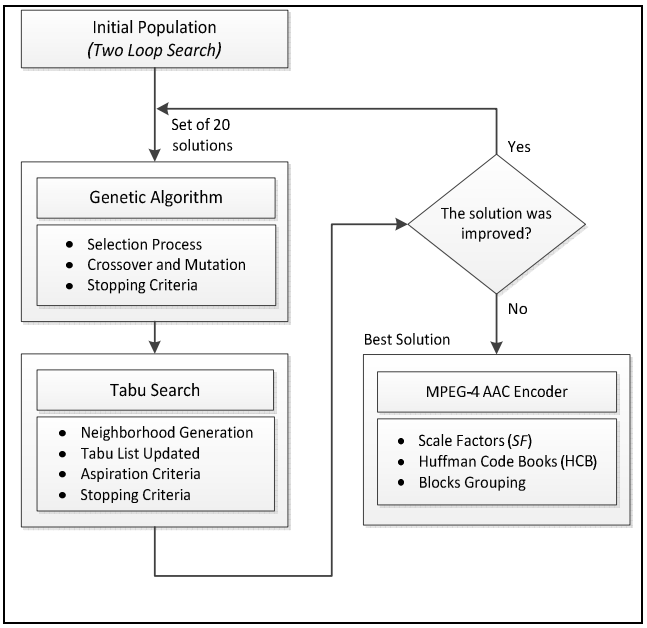


Figure 2. Hybrid architecture of the simulator.

Step 2: Utilizing the initial population, the GA generates new individuals, through crossover operator, and diversifies the population, through mutation operator. The process of selection by tournament, with a probability of 0.8 for the fittest, was used to identify individuals that will suffer crossover process. After the stop criterion is satisfied, the best solution obtained in the GA process is assumed as the initial solution of the TS. The adopted stopping criterion for TS process was a predefined number of iterations without improvement in OF value.

Step 3: From the initial solution, the TS generates neighbor solutions in an iterative process. The best solutions obtained in this process are stored in a new population, which is used as the input of the GA. TS is executed until the stop criterion is reached.

When the algorithm is finalized, the best solution obtained is passed to the MPEG-4 AAC encoder. From the data generated by the simulator, the frame encoding can be performed. The data of the ANMR are archived, for every frame, in order to allow the evaluation of the perceptual quality of the coded audio signal.

VI. COMPUTATIONAL EXPERIMENTS

The audio encoder and the simulator were programmed using the ANSI C language. Experiments and validation tests were performed through audio files. In this way, audio files in WAV format were used as input of encoder, which generates compressed audio files in AAC extension in its output. For the coding experiments a set of audio test was necessary, covering different types of music and sounds. Thus, based in [3][13][14][15], an audio files library was prepared to be used in the experiments of this work. The list of sounds composing the audio library test is shown in Table I. All WAV files listed have one audio channel with a sampling frequency of 44100Hz, and each sample is represented by 16 bits. To validate the developed computational model, experiments were performed in order to identify the influence of the three decision variables addressed in this study on perceptual quality metrics ANMR and bit rate. Through these experiments, the individual behavior of each one of the decision variables was studied. Thus, the value of only one of the variables was changed, while the values of other variables were held constant, totaling three validation experiments. As said before, the implementation was developed in ANSI C language, and it was inserted in the code of the Ffmpeg reference encoder, and was run on an Intel® Core™ i3 computer.

TABLE I. TEST AUDIO FILES

Index	Sound Type	Description	Duration
1	Drums	DG Samples – Rock beat Drums 02	8,15s
2	Blues	All Blues – Kora Jazz Band	15.03s
3	Bass	The Clairvoyant – Iron Maiden	9.50s
4	Choral	Symphony N^0 9 – Beethoven	10.03s
5	Crash China	DG Samples – Crash China	3.16s
6	Harpsichord	Oeuvres Pour Clavecin – François Couperin	10.02s
7	Narration	DG Samples – Vocals Shout 120	10.85s
8	Orchestra	Symphony N^0 9 – Beethoven	15.40s
9	Rock	It´s my life – Bom Jovi	15.07s
10	Sax	DG Samples – Sax Riff 128 C	8.24s

In the first experiment, to identify the individual behavior of SF, all frames are intentionally encoded as long blocks, and the fifth HCB was used, so that both variables would not influence on the measure of perceptual quality and bit rate. Following, the SF values were changed within a specific range [7] and [8], and its relation to the bit rate and perceptual quality is presented in Figure 3.

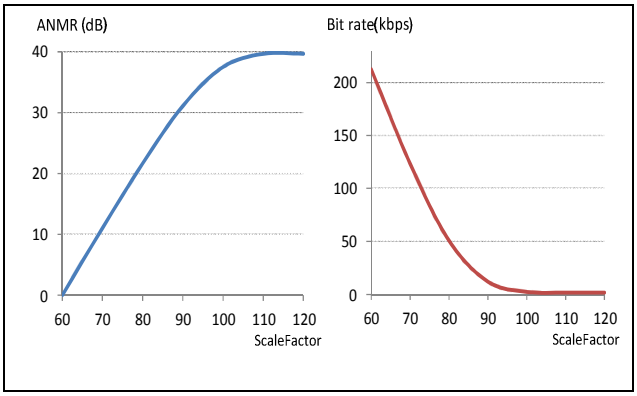


Figure 3. Scale Factor *vs* ANMR and bit rate.

The graphical analysis of the Figure 3 identifies that the smaller the value of SF, lower distortion is present in audio signal, i.e., greater the perceptual quality, and greater the number of bits required.

Similarly, the second validation experiment, all frames were also configured as long blocks, and the SF values were kept constant with a value 90. The values of HCB were changed in twelve possible values, and the results can be seen in Figure 4.

Through a graphical analysis of the Figure 4, the ranges of Largest Absolute Value (LAV) [7] and [8], from HCB are clearly identified, due the distortion of the audio signal does not vary for HCB of even values of LAV.

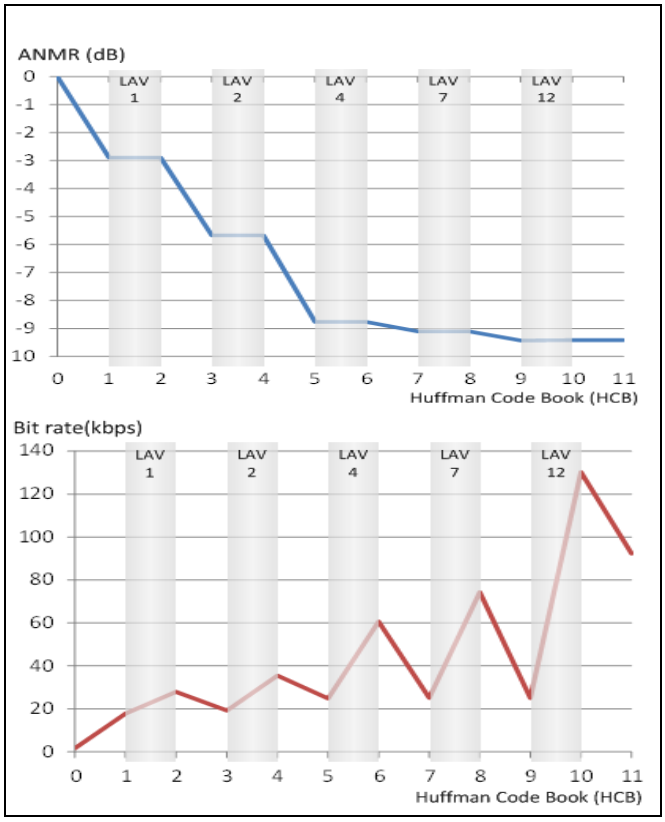


Figure 4. Huffman code book *vs* ANMR and bit rate

However, the bit rate is changed for HCB of even values of LAV, and increases as the value of LAV also increases [7]. The third experiment was performed to identify the influence of the grouping of blocks into sections, configured as a frame short block, on the metric ANMR and bit rate. Thus, all frames have been intentionally configured as short block, and the SF and HCB values were fixed at 90 and 5, respectively. In this experiment, combinations of seven sections were used, with the number of sections incremented from two to eight. It is observed that with the increased number of sections within one frame, the distortion value was reduced, and bit rate value was increased. This behavior follows the standard documentation [7] and [8] and is shown in Figure 5.

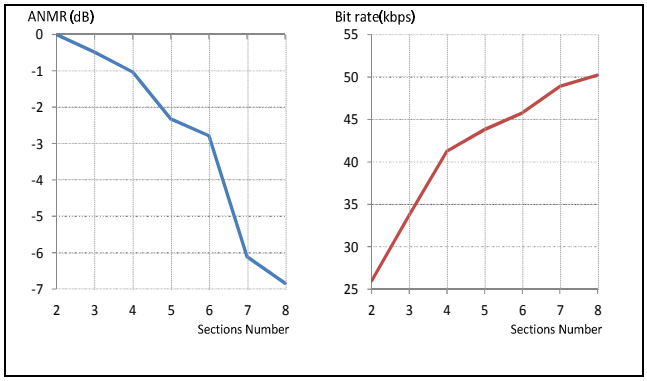


Figure 5. Huffman code book *vs* ANMR and bit rate

The dynamics observed in the three validation experiments, using the developed simulator, represent the behavior of the MPEG-4 AAC encoder [7] and [8] faithfully. Thus, the developed simulator is validated.

To set the calibration parameters of GA and TS, extensive experiments were conducted to identify which parameters combination that produced better values of OF. In GA the following values were set: mutation probability 0.9, crossover probability 0.8, and maximum number of generations 50. In the TS algorithm the value 200 was used for define the number of interactions without improvement in the FO value, utilized in the stop criterion, and the size of Tabu List. To evaluate the performance of the simulator developed for the MPEG-4 AAC encoder, the ten test audio files were encoded in four predetermined bit rates: 48kbps, 116kbps, 184kbps and 250kbps. Thus, it was possible to verify if the simulator model has produced better results regarding perceptual audio quality when compared to the reference encoder, using the TLS technique and Hybrid Algorithm (HA). The experimental results can be analyzed from Table II. Due to the stochastic behavior of metaheuristics, 50 encoding experiments for every audio file were executed, so that the results can be expressed by its mean and standard deviation values. For all audio files test, the HA had provided lower results for ANMR, i.e., a superior perceptual quality if compared to those obtained with TLS technique from reference encoder. The average execution time, per frame, for the rate of 48kbps was 50ms, 116kbps was 150ms, 184kbps was 300ms, and for 250kbps was 400ms. Table II presents the average of the attenuation of distortion and the respective standard deviation for the ten audio files tested, obtained by hybrid algorithm in relation to the process TLS, from reference MPEG-4 AAC encoder. The average values and the respective standard deviations were obtained from 50 runs of the simulator for all values shown in Table II.

 ISBN: 978-1-61208-398-8

TABLE II. RESULTS OF THE EXPERIMENTS.

File Index	Distortion (ANMR)							
	48kbps		116kbps		184kbps		250kbps	
	TLS	HA	TLS	HA	TLS	HA	TLS	HA
1	3.467	2.485 σ = 93	685	404 σ = 49	116	94 σ = 7	21	14 σ = 1
2	9.541	3.778 σ = 86	1.391	422 σ = 10	171	82 σ = 3	48	16 σ = 0,497
3	275	143 σ = 37	53	2 σ = 0,102	3	0,416 σ = 0,041	2	0,106 σ = 0,002
4	1.519	1.356 σ = 69	115	65 σ = 10	24	10 σ = 0,233	6	2 σ = 0,59
5	9.306	6.197 σ = 172	681	563 σ = 19	191	92 σ = 8	39	15 σ = 1
6	5.131	2.708 σ = 83	682	266 σ = 24	81	52 σ = 2	18	10 σ = 0,26
7	410	117 σ = 17	89	13 σ = 1	4	3 σ = 0,487	3	0,59 σ = 0,027
8	934	740 σ = 72	114	44 σ = 7	23	8 σ = 0,226	4	2 σ = 0,034
9	55.476	27.327 σ = 177	11.501	3.252 σ = 55	1.084	551 σ = 7	354	102 σ = 2
10	1.401	1.130 σ = 76	312	107 σ = 13	19	11 σ = 1	4	2 σ = 0,189

In Table III, it can be observed the gain (dB) of perceptual quality of the solutions found by the simulator,

TABLE III. HYBRID ALGORITHM *VS* TWO LOOP SEARCH (TLS).

ANMR (dB)	Bit rate			
	48 kbps	116 kbps	184 kbps	250 kbps
μ	-2,38	-5,17	-3,27	-4,92
σ	1,55	3,79	2,18	3,16

when compared to the solutions of reference MPEG-4 AAC encoder. The negative values represent a decrease of values obtained by the ANMR metric.

VII. CONCLUSIONS

This work has proposed a simulator that uses the GA in combination with the TS, which was developed to simulate the process of solve the AAC Encoding Problem. Currently, this problem is solved using the technique of the reference encoder MPEG-4 AAC, known as TLS.

The simulator was validated through extensive experiments, showing the dynamic behavior as expressed in the MPEG-4 AAC standard [9]. The experimental results showed a significant improvement in perceptual audio quality achieved by the HA proposed, when compared to the TLS technique from reference encoder. In all test runs, for different files and different bit rates, the simulator has produced the best results. For higher bit rates, processing time can be significant.

REFERENCES

[1] C. Bauer and M. Vinton,"Joint Optimization of Scale Factors and Huffman Code Books for MPEG-4 AAC," IEEE Transactions on Signal Processing, vol. 54, no. 1, Jan. 2006, pp. 177-189, doi: 10.1109/TSP.2005.861090.

[2] A. Aggarwal, S. L. Regunathan and K. Rose, "Near - Optimal Selection of Encoding Parametersf for Audio Coding," Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May, 2001, vol. 5, pp. 3269-3272, doi: 10.1109/ICASSP.2001.940356.

[3] C. Bauer, "The Optimal Choice of Encoding Parameters for MPEG4 AAC Streamed over Wireless Networks," Proc. 1st ACM Workshop on Wireless Multimedia Networking and Performance Modeling (WMuNep'05), Oct. 2005, pp. 93-100, doi: 10.1145/1089737.1089753.

[4] F. Glover, " Future Paths for Integer Programming And Links to Artificial Intelligence," Computers & Operations Research, May 1986, vol. 13, no. 5, pp. 533-549, doi:10.1016/0305-0548(86)90048-1.

[5] J. H. Holland, Adaptation in Natural and Artificial Systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, Michigan, 1st edition, 1975.

[6] C. Bauer, M. Fellersand and G. Davidson, "Multidimensional Optimization of MPEG-4 AAC Encoding." Proc. International Conference on Acoustics, Speech, and signal Processing (ICASSP), May 2006, vol. 5, pp. 69-72, doi: 10.1109/ICASSP.2006.1661214.

[7] ISO/IEC,14496-3: Information Technology - Coding of Audio Visual Objects - Part 3: Audio, Geneva, 2005.

[8] ISO/IEC, 13818-7: Information technology - Generic Coding of Moving Pictures and Associated Audio Information : Part 7 - Advanced Audio Coding (AAC), Geneva, 2004.

[9] H. Fastl and E. Zwicker, Psychoacoustics: Facts and Models, Springer, New York, 3rd ed., 2007.

[10] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE Journal on Selected Areas in Communications, vol. 6, no. 2, Mar. 1988, pp. 314-323, doi: 10.1109/49.608.

[11] M. T. Ali and M. Saleem, "Improved Audio Quality at 48 Kbits/s for MPEG-4 AAC," Proc. Second International Conference on Electrical Engineering, Mar. 2008, pp. 1-8, doi: 10.1109/ICEE.2008.4553921.

[12] B. Wang, J. Zhang, J.,Y. Yao and L. Xie, "A New Bit-allocation Algorithm for AAC Encoder Based on Linear Prediction," Proc. 11ª International Conference on Communication Technology Proceedings, Nov. 2008, pp. 726-729, doi: 10.1109/ICCT.2008.4716222 .

[13] S. Wu and X. Qiu, "An Bit Allocation Method Based Rate-Distortion Control Algorithm for MPEG-4 Advanced Audio Coding," Proc. International Conference on Audio, Language and Image Processing (ICALIP2008), Jul. 2008, pp. 237-241, doi: 10.1109/ICALIP.2008.4590241.

[14] O. Derrien, P. Duhamel, M. Charbit and G. Richard," A New Quantization Optimization Algorithm for the MPEG Advanced Audio Coder Using a Statistical Subband Model of the Quantization Noise," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, Jun. 2006, pp. 1328-1339, doi: 10.1109/TSA.2005.858041.

[15] N. C. Singh, "Measuring the Complexity of Sound," Pramana - Journal of Physics, vol. 77, no. 5, Nov. 2011, pp. 811-816, doi: 101007/s12043-011-0188-y.

 ISBN: 978-1-61208-398-8

Computational Clusters Efficient Parallel Data Transmission Paradigm

Mahdi Qasim Mohammed
Mohammed M. Azeez
Mustafa Aljshamee
Distributed high performance computing institute
University of Rostock
Rostock, Germany
mahdi.mohammed@uni-rostock.de
mohammed.azeez1983@yahoo.com
mustafa.aljshamee@uni-rostock.de

Abbas Malekpour
Peter Luksch
Distributed high performance computing institute
University of Rostock
Rostock, Germany
abbas.malekpour@uni-rostock.de
peter.luksch@uni-rostock.de

***Abstract*— Message Passing Interface (MPI) is the most popular and widespread model used for distributed parallel programming in high performance computing systems. Stream Control Transmission Protocol (SCTP) is a standard transport layer protocol. It supports a lot of features not supported by transmission control protocol (TCP), recently the standard transport layer protocol used by MPI, which make SCTP a promising solution to be used as MPI under layer protocol. In this work, Multi-streaming and Multi-homing, the most powerful features supported by SCTP, are used to implement parallel data transmission over computational networks using all the available paths and physical interface cards. For this purpose, an application programming interface API had been designed and used.**

Keywords: Message passing interface (MPI),Stream control transmission protocol (SCTP) .

I. INTRODUCTION

The tremendous increase in the utilization of computer networks in different applications and purposes and the raise in performance in networks' features like security, reliability, bandwidth and throughput in the last decades encouraged the computer programmers to benefit these parameters to establish distributed and high performance computing systems. High performance computing systems create an adequate environment for executing programs in parallel and getting the advantages of the computer resources distributed over the network. Message Passing Interface MPI is a portable model to build distributed systems and create an efficient environment for different scientific applications. Stream Control Transmission protocol (SCTP) is a modern transport layer protocol with new and efficient features that are not supported by TCP. SCTP combines the best features of TCP and User Datagram Protocol (UDP). It is message oriented protocol like UDP and provides effective algorithms for connection management like congestion control and flow control [2] [3].

In this work, we used SCTP as a transport layer protocol. We focused on using Multi-streaming and Multi-homing features in MPI applications and test the feasibility of using the aforementioned features on the performance of network throughput. An application programming interface API for MPI applications had been designed and implemented for this purpose[1][3][7][15].

The structure of this paper will be as follows. After introduction SCTP protocol and its features multi-homing and multi-streaming have been described then we discuss the most related projects and works to this paper. Then we describe and explain our work and different testing scenarios and the testing environment. Then, we discuss the results and the difference between the expected and the real results. Finally, the concluding section summarizes the article.

II. STREAM CONTROL TRANSMISSION PROTOCOL

SCTP is a transport layer protocols and combines the best features of the other transport layer protocols (TCP and UDP). It is connection-oriented, general purpose, reliable, message-oriented protocol. SCTP supports Multi-homing and Multi-streaming which represent the most interested features in modern Internet networks. Figure 1 shows SCTP position in Internet layer protocols [3].

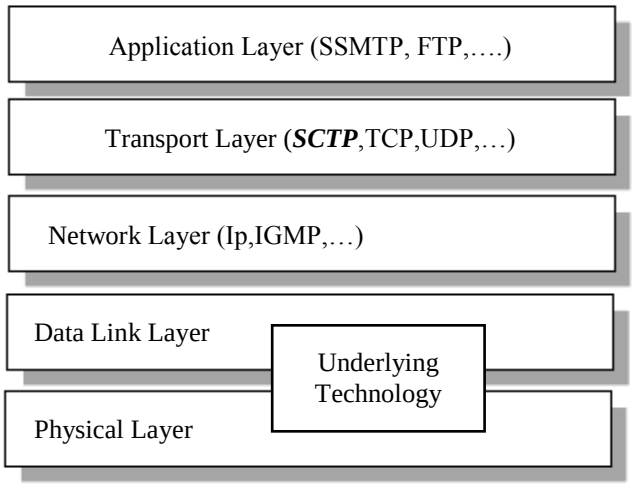


Figure 1. TCP/IP Protocol.

A. *Multi-homing*

In computer networks, each endpoint uses one IP address for data exchanging even when it has more than one IP address and connected to more than one network. This case represents a weak point in performance of network communication especially with the revolutionary growth in hardware industry. Currently, each endpoint contains more than one network interface card, which means, it is possible to be connected to more than one network simultaneously. SCTP solves this problem by supporting Multi-homing feature. The communicating endpoints can exchange data via multi-paths using multiple IP addresses. Association (connection between SCTP endpoints) chooses one IP address and its path at each endpoint to create primary path for data communication while the other addresses and paths used as a backup. If the primary path suffered from any obstacle, the other paths will be used for data transmission. Current SCTP does not support concurrent data transmission over all available paths [1] [3].

B. *Multi-streaming*

The connection between SCTP endpoints is called association. SCTP supports Multi-streaming features in its endpoints association. Multi-streaming is one of the most powerful and important features of SCTP. TCP is a single stream protocol and the drawback in single stream protocols is that any blocking to this single stream hinders the transmission of data between communicating endpoints. This problem had been exceeded in SCTP Multi-streaming feature by creating multi streams for data transmission between the endpoints. Each stream is independent from other streams. If one of the streams faces any problems, data transmission continues in the other streams without any side effects.. For example, if three streams existed between two SCTP endpoints transmitting data and a data loss occurred in stream number three, only this stream needs to wait for the retransmission operation while the transmission over stream one and stream two will not be affected because SCTP does not need to warn about consecutive data delivery over all streams; just stream three should wait for retransmission of its lost packets [1][3][4].

III. RELATED WORK

Dreibholz [1] has investigated the property of parallel data transfer by using SCTP's extension CMT. Tests and analysis showed there are no big differences in the performance when using standard SCTP and SCTP-CMT. Also, the author described three drawbacks. Firstly, unordered delivery of data at receiver side. Secondly, the unordered data received by application layer at the receiver side and this cause a problem for the applications that needs ordered data. The third problem resulted from the unsuitable resources allocation for different application's messages requirements. Kamal et al. [9] from the earliest projects suggested to use SCTP for MPI and used special module, request progressing interface RPI, to provide MPI's processes communication. Three methods were suggested and implemented to develop this RPI module to use SCTP instead of TCP. Iyengar et al. [10] Suggested utilization Multi-homing feature to implement concurrent multi path transfer of data. Three problems had been discovered as drawbacks associated to simultaneous transfer of data over multi paths, fast retransmission of already transmitted chunks at the sender side, reduced congestion window growth at the sender side and increasing of data traffic in the return path and solution to each one had been suggested. Penoff et al. [11] studied the feasibility of using Multi-homing and Multi-streaming features in MPI applications in computers clusters. They studied the effects of multi networks' configuration, messages scheduling, fault tolerance and congestion control on the cluster communication. Becke et al. [12] in their work developed and improved two load sharing transport layer protocols, SCTP-CMT and multipath TCP (MPTCP), and they compared the performance of those protocols by testing them over two networks, local network lab and real network between Germany and China over the internet. They discovered two important features that affects the performance of multi path networks, congestion control and path and resources management. Also, they suggested solutions depending on resource pooling to solve the problem of an inadequate use of the network's resources and proposed techniques to solve congestion control problem by splitting the congestion window to group of windows each one associated to one of the communicating paths.

IV. CMT FOR MPI: DESIGN AND IMPLEMENTATION

Recently, computational resources especially central processing units (CPUs) which are the main modules in the computational clusters became very high speed and performance efficient. Most of the current central processing units CPUs based on multi-core processors architectures with multi-level cache memory. In addition, some nodes in the computational clusters consists accelerators to support computational operations executed by CPUs. Computational accelerators include different devices and architectures like graphical processing units GPUs, field programmable gate arrays FPGAs. All these devices make cluster nodes massive computational devices, but on the other hand make the communication between these nodes the bottle neck of the total cluster performance. Efficient execution devices with unsuitable network communication can dramatically hinder computations achievement because this disparity in the performance. One of the good solutions to increase network communications is the use of multi-streaming communication with multi-homing feature. This solution utilizes all network interface cards available in cluster nodes to implement parallel data communication. This solution is the ideal one especially, for the currently established and in use clusters, because the replacement of

 ISBN: 978-1-61208-398-8

the used network's infrastructure, topology and devices with a new technologies and infrastructures like infiniband is a hard and expensive solution. Also, increasing network throughput can decrease energy consumed by the clusters and this is a great benefit in this time. In this work, we designed an application programming interface API to utilize Multi-streaming and Multi-homing for concurrent multipath data transfer implementation. For this reason, multiple streams established between SCTP endpoints using independent networks and IP addresses for simultaneous data exchanging. These paths are disjoint and related to independent network.

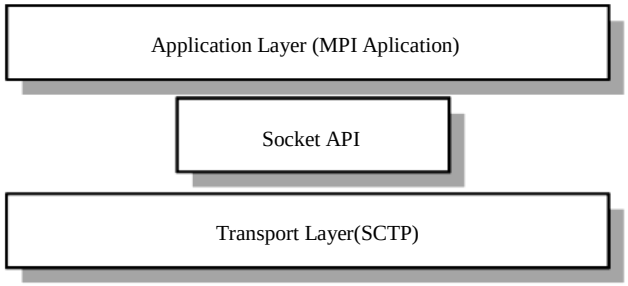


Figure 2. CMT Implementation in the Internet network Layers.

Standard SCTP establishes multiple paths between the communicated endpoints and uses one of these paths for data transmission and others as backup path. We used concurrent data transfer over multipath at the socket API layer, connecting socket between application and transport layers, not at the transport layer like aforementioned projects as shown in Figure 2. As mentioned before, standard SCTP uses one path as a primary one for data exchanging and the rest for backup but this API extends this feature to utilize all the established paths in data transmission.

The designed and developed API establishes independent transmission streams over different networks via network interface cards. Each message comes from the application layer, which is the upper layer, will be received by this API and then will be allocated to one of the established communication paths by implementing a Round-Robin scheduling algorithm to distribute the messages on the communication paths. This API receives user messages from application layer and distributes these messages over the communicating streams and start concurrent data transmission over the established paths. In this design each message that is received from application layer is allocated to specific resources and sent over independent network address interfaced to different networks.

Socket API is the middleware between the transport and application layers and able to utilize the resources of both of them and realizes the best usage of these resources. Working at this layer offers information on each type of messages created by different applications to allocate the required resources in dependence on the application's messages requirements making a better data exchanging performance. The designed API has been tested with four endpoints cluster with different network topologies. In the first test each endpoint interfaced to one network (single homed) while in the second test each one endpoint interfaced to two networks (dual homed). In the last test, endpoints interfaced to three networks (triple homed).

The testing environment had been built using OMNeT++ simulation environment and its framework Inet as shown in Figures 3-5. Each endpoint had been represented by an instance of the StandardHost module which is a module in OMNeT++'s INET framework. These endpoints used SCTP as a transport layer protocol. Each two endpoints had been connected by an instance of the Router module. The connection channels between the elements of the network are instances of the module DatarateChannel. In this module the data rate and the delay time of each channel could be changed to test the simulated network in different scenarios. The network was configured by using an instance of the module IPv4NetworkConfigurator to provide each endpoint with the appropriate IP address. MPI-NeTSim module was used to interface MPI applications to the simulated networks that was built by OMNeT++.

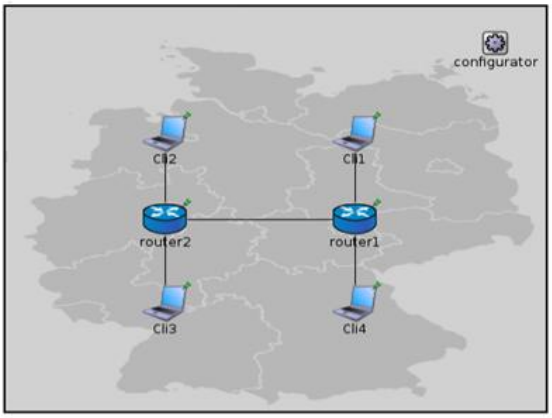


Figure 3. Single homed network.

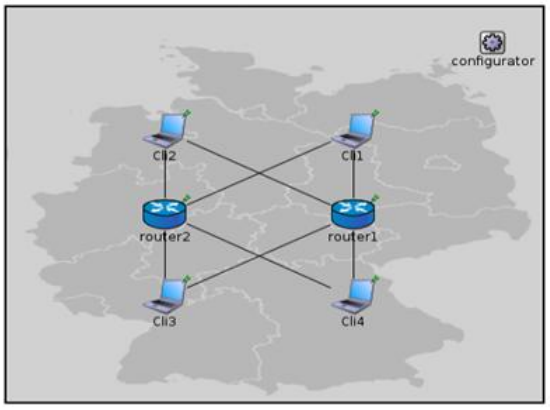


Figure 4. Dual homed network.

Figures 5-6 and Tables 1-2 show the performance of dual and triple homed cluster versus single homed cluster. We can clearly notice that the multi-homed clusters realize a better performance than the single homed and achieve

higher data exchange rate leading to a faster execution time and better utilization of network resources.

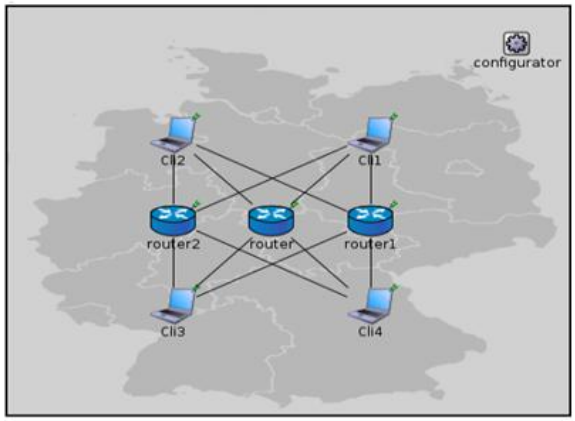


Figure 5. Triple homed network.

TABLE I. SINGLE AND DUAL HOMED NETWORK PERFORMANCE

Data Size MB	*Execution Time mSec*	*Execution Time mSec*
20	2.286814	1.351112
24	4.464083	3.578660
28	8.456284	6.659053
32	13.645028	10.097254
36	19.722385	14.629627
40	30.331633	19.855265

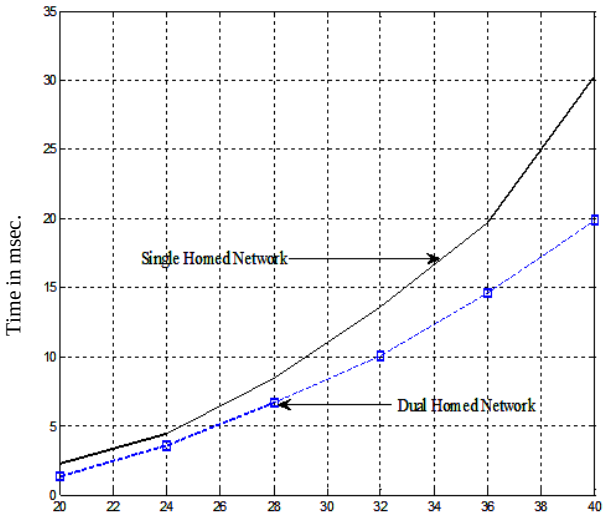


Figure 6. Singel versus dual homed network performance.

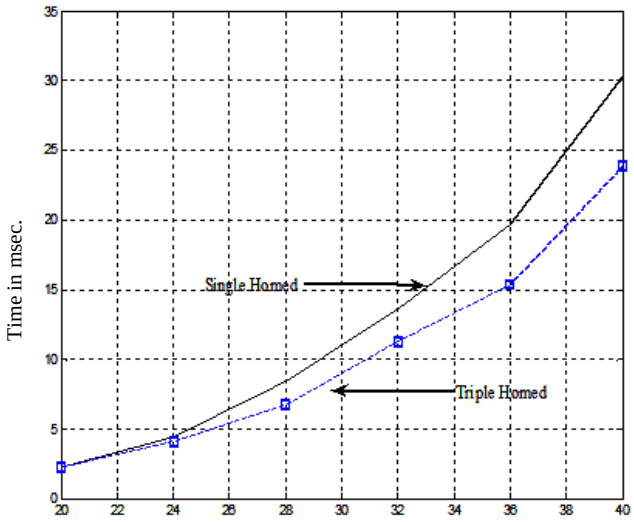


Figure 7. Singel versus Triple homed network performance.

TABLE II. SINGLE AND TRIPLE HOMED NETWORK PERFORMANCE

Data Size MB	*Execution Time mSec*	*Execution Time mSec*
20	2.286814	2.236239
24	4.464083	4.111785
28	8.456284	6.756544
32	13.645028	11.328705
36	19.722385	15.363163
40	30.331633	23.849022

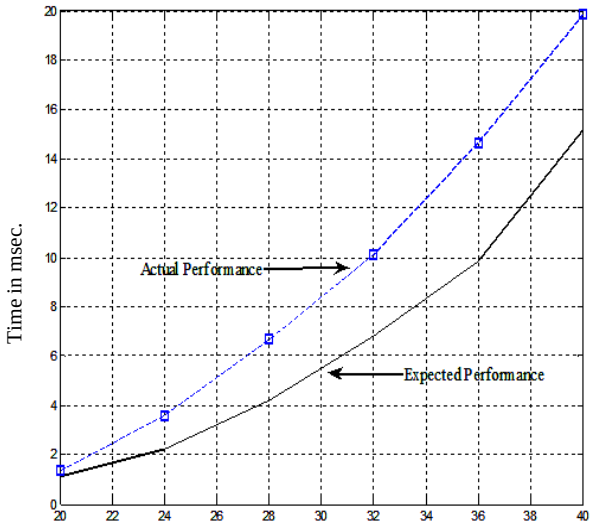


Figure 8. Real and expected performance of dual homed network.

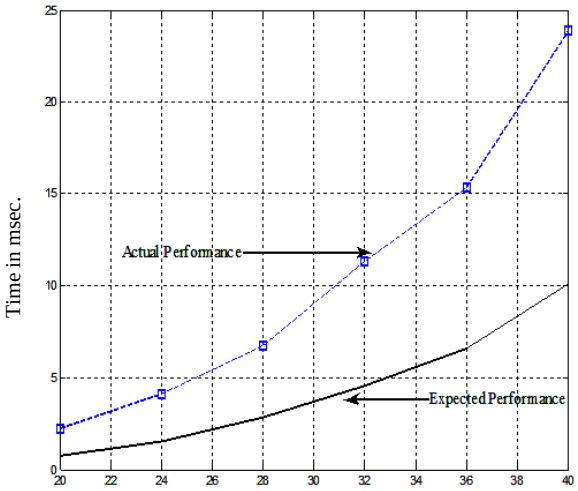


Figure 9. Real and expected performance of triple homed network.

At the same time, figures 8-9 show that we did not acquire the expected improvement in the performance because of increasing number of paths between endpoints should increase network throughput and decrease the time required for data exchanging but concurrent multipath transfer of data is associated to unnecessary fast retransmission of delayed data packets or few updates of sender congestion window which decrease the growth of this window and network throughput in addition to inadequate resources allocation for different messages requirements. These problems become more severe when increasing number of networks interfaced to cluster nodes and hinder the performance and cause these drawbacks in the designed API performance.

V. CONCLUSION

In this work, an application programming interface API has been developed to investigate the use of concurrent multipath data transfer in multi-homed computer clusters. A faster data transfer for MPI applications has been realized but stills lower than ideal performance that expected. The results showed also that the API performance decreases when the number of networks interfaced to endpoints increased.

The principal of concurrent multi path data transfer is expected to play a main role in different scientific computing applications to improve its performance. Also concurrent multipath transfer of data can play a main and essential role in high performance cloud computing like Hoopoe which is a GPGPU-based cloud computing, which is classified as a faster than real time infrastructure and has been used in different scientific applications [16] and needs a fast communication environment to prepare and transmits data with high data rates to integrates the work of this cloud computing. For this reason, the concurrent multipath data transfer will be the best solution to provide a required communication environment.

The most important point has been discovered in this work is that the increasing number of paths for data communication is not always the right solution to improve the performance and throughput of the network because concurrent data transfer is associated with problems of retransmission lost or delayed data packets and large congestion windows. As a result, to improve concurrent data transfer over multipath and different networks definitely it is needed to improve the total mechanisms of data packets management, retransmission policy and congestion window growth in SCTP protocol.

REFERENCES

[1] T. Dreibholz, “Evaluation and Optimisation of Multi-Path Transport using the Stream Control Transmission Protocol”, 2012.

[2] H. Kamal, “SCTP-Based Middleware for MPI”, 2005.

[3] B. A. Forouzan , “TCP/IP Protocol Suite”, 2010.

[4] http://www.ibm.com/developerworks/library/l-sctp/

[5] F. Perotto, C. Casetti and G. Galante, “SCTP-based Transport Protocols for Concurrent Multipath Transfer”, WCNC 2007, pp 2971-2976.

[6] A. Varga, OMNeT++ User Manual Version 4.3, 2011.

[7] M. Snir, S. Otto, S. Huss-Lederman, D. Walker and J. Dongarra, “MPI the complete reference”, 1996.

[8] http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101743820020001000007.

[9] H. Kamal, B. Penoff and A. Wagner, “SCTP-based Middleware for MPI in Wide Area Networks”,(CNCR05) 0-7695-2333-1/05 ,2005.

[10] J. R. Iyengar, K. C. Shah and P. D. Amer, “Concurrent Multipath Transfer Using SCTP Multihoming”.

[11] B. Penoff, M. Tsai1, J. Iyengar and A. Wagner, “Using CMT in SCTP-Based MPI to Exploit Multiple Interfaces in Cluster Nodes”, EuroPVM/MPI 2007, LNCS 4757, pp. 204–212, 2007.

[12] M. Becke, H. Adhari, E. P. Rathgeb, F. Fa, X. Yang and X. Zhou, “Comparison of Multipath TCP and CMT-SCTPbased on Intercontinental Measurements”.

[13] http://artemis.wszib.edu.pl/~mradecki/prir/lab2/.

[14] B. Penoff, A. Wagner, M. Tuxen and I. Rungeler, “MPI-NeTSim: A network simulation module for MPI”.

[15] J. R. Iyengar, P.l D. Amer and R. Stewart, “Concurrent Multipath Transfer Using SCTP Multihoming Over Independent End-to-End Paths”, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 14, NO. 5, OCTOBER, pp 951-964, 2006.

[16] http://www.cass-hpc.com/solutions/hoopoe.

 ISBN: 978-1-61208-398-8

MannaSim: A NS-2 extension to simulate Wireless Sensor Network

Rodolfo Miranda Pereira, Linnyer Beatrys Ruiz
Department of Informatics, State University of Maringa
Maringa, Parana, Brazil
e-mail: rodolfomp123@gmail.com, linnyer@gmail.com

Maria Luisa Amarante Ghizoni
Engineer School, Federal University of Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
e-mail: malu.ghi@gmail.com

***Abstract*—A Wireless Sensor Network is a special kind of ad hoc network characterized by a large amount of nodes distributed over a real world environment, in order to monitor and send relevant data to an access point. In this paper, we present a framework for wireless sensors network simulation that was implemented as an extension for Network Simulator 2. This framework provides a way to configure the environment of the simulation by choosing parameters like the topology of the network, kind of data dissemination, type of routing protocol, hardware capacity and initial energy power of the sensor nodes. In order to demonstrate MannaSim's results, we also present a wireless sensor network scenario in which the proposed tool was used to simulate. Through the results provided by MannaSim, it was possible to choose the best configuration for the proposed scenario.**

Keywords–Wireless Sensor Network; Network Simulator (NS-2); Simulation Framework.

I. INTRODUCTION

Wireless Sensors Networks (WSN) are a class of distributed systems that recently is being target of a lot of research [1]. In a WSN, these data are captured by the sensor nodes in the environment that they are involved. Despite their variety, all WSNs have certain fundamental features in common. The essential is that they are embedded in the "real world". A WSN basically consists of sensor nodes deployed over a geographical area for monitoring physical phenomena like temperature, sound, vibrations, humidity, seismic events, chemical elements, and others. For this reason, WSN have many applications like smart environments in homes [2], buildings [3], transportation [4], disaster prevention [5] and even in the pharmaceutical field [6]. WSN have captured the attention and imagination of many researchers, encompassing a broad spectrum of ideas.

In the field of computer networks, in general, the simulation allows the evaluation of scenarios with a small cost and time compared to experiments in physical environments. This avoids unnecessary costs assembling real networks, and allows desired comparisons, improving decision-making power in the choice of the network parameters. In the case of WSNs, the simulation is even more advantageous, because it can be composed by a lot of sensor nodes, increasing the costs of this kind of network. Beyond this, different applications requires different types of sensor nodes, so the simulations can assist in decisions about the most appropriate sensor node to be utilized. Finally, changing and testing the configurations of the WSN certainly can help to understand the relation between the parameters.

Although simulation can bring many benefits, the choice of the wrong simulator may not result in satisfactory results. Simulations are very dependent of the functional model developed, and the model should optimally represent the environment simulated. MannaSim was made as an extension for the consecrated Network Simulator (NS-2) [7]. The framework consists of a set of base classes for the simulation of WSN, which may be specialized by users, and these classes extend the core of the NS-2 Simulator. MannaSim was built on existing model for WSN [8] and it is configurable in terms of sensing platform. As NS-2 Simulator, MannaSim is open source, thus the user can adapt the code if necessary.

MannaSim allows the user to configure detailed scenarios for simulations. Setting compositional requirements of the network (number of nodes, node type, density, dissemination type) and its organization (flat or hierarchical) MannaSim can accurately model different sensor nodes and applications while providing a versatile testbed for algorithms and protocols. After the simulation, MannaSim generate results, like the power level remaining in the components, the number and type of errors in the simulation, the average events division and the operation of the network in its lifetime. This kind of information can be analyzed and taken into account choosing the best configuration for the proposed WSN.

The rest of this paper is organized as follows: Section II presents some related works. The MannaSim extension is, as well as its design and the Script Generator Tool, presented in Section III. Section IV shows a simulation example, with the scenario description and the results of the simulations. Finally, in Section V, the conclusions and future works are presented.

II. RELATED WORK

A large number of WSN simulators has been published with different simulation outbreaks, requiring that developers are aware of these simulators so they can make the best choice of simulation in their projects. In the following, a review of some simulators already published will be shown.

Sensor Network Simulator and Emulator (SENSE) [9] is a simulation tool that is based on a component-oriented methodology that intends to promote extensibility and reusability to the maximum degree. It was designed in order to attend 3 kinds of users: high-level users, network builders and component designers. SENSE also proposes to be as efficient as NS-2 and as scalable as possible in its simulation. However, compared to the MannaSim framework, SENSE still lacks a comprehensive set of models and a wide variety of configuration templates that are required for wireless sensor network simulations.

SensorSim [10], which also has been built on top of NS-2, is a simulation framework for sensor networks. It provides sensor channel models, energy consumers, lightweight protocol stacks for wireless micro sensors, scenario generation and hybrid simulation. The sensor channel models the dynamic interaction between the sensor nodes and the physical environment.

 ISBN: 978-1-61208-398-8

At each node, energy consumers are said to control the power in multiple modes to efficiently use the power and prolong the nodes lifetime. When compared, MannaSim has much more scenarios configurations than SensorSim. Besides, SensorSim is no longer developed, therefore, no more available.

Another WSN simulator that we can quote is Avrora [11]. This framework is a cycle-accurate instruction level WSN simulator, that uses an event-queue model that allows improved interpreter performance and enables an essential sleep optimization. Its main goal is to provide the user the conditions to validate time-dependent properties of a large-scale WSN. Although Avrora intents to simulate the sensor nodes behavior at instruction level, it does not deal with the fact that nodes may run at slightly different clock frequencies over time due to manufacturing tolerances, temperature and battery performance. Compared to Avrora, MannaSim has a different purpose. While Avrora intends to provide a instruction level simulation, MannaSim proposes to give the user an event-based overview of the nodes communication.

Castalia [12] is a WSN simulator built on top of OMNeT++. Castalia features an accurate channel/radio model detailing radio behaviour. It also features a flexible physical process model, taking into account issues such as clock drift, sensor bias, sensor energy consumption, CPU energy consumption, and monitors resources such as memory usage and CPU time. Castalia's goal is to provide the user conditions to test algorithms and protocols in a wireless channel and radio model, with a realistic node behaviour relating to access of the radio. Thus, while Castalia focuses on simulating a radio model behaviour, MannaSim is aimed to provide a validation of the impacts of different settings and compositions on the errors and energy consumption of the sensor nodes.

NetTopo [13] is a framework for WSN simulation that has a visualization function to assist the investigation of algorithms in WSNs. NetTopo provides a common virtual WSN for the purpose of interaction between sensor devices and simulated virtual nodes. It allows users to define a large number of initial parameters for sensor nodes like residential energy, transmission bandwidth and radio radius. Users can also define and extend the internal processing behavior of sensor nodes like energy consumption and bandwidth management. For the visualization module, it works as a plug-in component to visualize testbed's connection status, topology and sensed data. Compared to MannaSim, NetTopo's goal is also different. NetTopo intends to assist the investigation of different algorithms impacts in a WSN, while MannaSim's goal is to investigate different compositional requirements impacts, like the network organization and number of nodes.

Table I shows a comparison between the main objectives and characteristics of MannaSim and others WSN simulation tools. In summary, it can be said that the main objective of MannaSim is to provide the user a first-order simulation for generic sensor nodes platforms in order to arrange an investigation of different network organizational and characteristics.

TABLE I. MANNASIM VS. OTHER SIMULATION TOOLS.

Framework	Characteristic	MannaSim
SENSE	Was designed in order to attend 3 kinds of users: high-level users, network builders and component designers, but still lacks a comprehensive set of models to configurate.	Was developed to first order validation users and has a lot of configurable parameters for the network composition.
SensorSim	Also has been built on top of NS-2, but is no longer developed.	Can give much more configurable parameters in the WSN, and is still under development.
Avrora	Simulate sensor nodes behavior at instruction level.	Gives the user a based on events overview of the nodes communication during network lifetime.
NetTopo	Its target is to investigate the impact of different algorithms in a WSN.	Its goal is to investigate the impacts of different network compositional requirements, such as organization and the number of nodes.
Castalia	Provides the user conditions to test algorithms and protocols in a realistic model of the nodes communication.	Proposes a first order validation in the chosen composition of the WSN.

III. THE MANNASIM FRAMEWORK

MannaSim is composed of two solutions: The Framework and the Script Generator Tool (SGT). The Framework is a module for WSN simulation based on the Network Simulator (NS-2) and SGT is a front-end for the creation of simulation scripts. MannaSim's home page [14] gives detailed informations about scenario configuration and also presents scenario examples. It is important to note that MannaSim is being developed under the GNU General Public License, in other words, it intends to guarantee the user freedom to share and change all versions of the code. In the following subsections, we present details about the framework.

A. On the MannaSim Design

MannaSim inherits the core features of the Network Simulator (NS-2), and builds up new features that include ability to use different protocol profiles for different WSN applications, different sensor parameters and distribution. The requirements of the network composition, like number of nodes, types of nodes, density, flat or hierarchical organization are different for each application. Thus, a WSN Simulator has to be flexible enough to attend this kinds of characteristics. The goal of MannaSim is to be flexible to make a detailed WSN simulation which can accurately model different sensor nodes and applications.

The first step taken in the implementation of the simulator was the implementation of a node specific to WSNs, the sensor node. Since NS-2 already possesses an object class that represents a mobile node with wireless communication capability, the new node was implemented extending the mobile nodes class. To this new node, new characteristics were added such as sensing and processing energy consumption, 'wake up' and 'sleep' functions and control of components usage state such as sensor devices and processor. A subclass of the existing energy model was also created; it implements a battery class that can be used to implement the different existing battery models. Next, specialized classes that describe the behavior of each node type found in a WSN were modeled and implemented. These behaviors were implemented in the application layer, since no restriction may be imposed to the user regarding the desired protocol stack. Thus, each developed class that models a node from MannaSim inherits from NS's application. Common-nodes, leader nodes and access points

 ISBN: 978-1-61208-398-8

were created also. Figure 1 shows the simplified class diagram of MannaSim.

Following the characteristics of a WSN, below we present how MannaSim's classes were designed to attend the possible features in a WSN.

- **Simulate different kinds of sensor devices:** Data collection in MannaSim is simulated through the generation of artificial informations. The class DataGenerator is the basis for the generation of information. Simply by extending it the user can simulate a variety of sensors devices. The artificial data collected must be encapsulated in a corresponding class. This class represents the data that will be disseminated by the sensor nodes in the WSN towards the AP.
- **Contemplate different sensing options:** MannaSim allows the continuous collection, periodically and on demand. The frequency which the data are generated by the inherited classes of DataGenerator models the different types of sensing. For networks that use scheduled or continuous collection, a timer (SensingTimer) is used. The demand network only performs the collection when a requisition is requested by the observer.
- **Contemplate different disseminating options:** MannaSim allows continuous, scheduled or on demand data dissemination, regardless of the chosen sensing type. For example, the network can collect data continuously, but spreads them periodically. For networks that utilize programmed dissemination, a timer (DisseminatingTimer class) is used. The demand network performs dissemination only if the observer sends a request. The requests are modeled by the class OnDemandData. Each request can contain multiple queries, which are instances of the class OnDemandParameter, it specifies the data of interest to the observer. WSNs with continued dissemination transmit data as soon as they are collected and processed. Messages of data that are sent to head nodes or to the AP, are modeled by the class SensedData, which is an implementation of the abstract class AppData API2 the standard NS-2.
- **Contemplate different processing options:** In MannaSim, all data collected pass through some kind of processing before being disseminated. The base class Processing serves as a starting point for the creation of specific types of processing for each application. The Processing of requests on demand are implemented in this class.
- **Allows the simulation of flat and hierarchical sensor networks:** The behavior of the sensors nodes was implemented in MannaSim as a protocol from the application layer in NS-2, using the inheritance from the Application class. The general behavior of a sensors is implemented in the class sensorBaseApp. This class have the basis for creating different types of behavior such as, for example, the head (class ClusterHeadApp) or the common (class CommonNodeApp). The creation of different behaviors allows modeling of hierarchical multilevel WSNs.
- **Allows the simulation of homogeneous and heterogeneous sensor networks:** Through the SensorNode class, inherited of the MobileNode class from NS-2, the MannaSim is able to create sensors nodes with different settings. Each sensor node has an object Battery class, a specialization class from EnergyModel of NS-2. This class defines a battery model for the sensors. If extended, new models of energy decay can be created in the simulations.
- **Simulate networks with one or more Access Points:** The MannaSim allows that a simulation of a WSN has one or more APs. The class AccessPointApp enables the communication of the network with the external observer. One or more nodes of the sensor network, or even a node that is not a sensor node, may contain this application and act as an AP.
- **Allows the utilization of different protocols:** Different routing protocols (specific or not to WSNs, single or multi-hop) can be used. The same applies to the transport and link layers. The use of protocols in these layers, in the simulations with the MannaSim, follows the same procedure for any NS-2 simulation. Two of the most popular routing protocols for WSNs (LEACH [15] and Directed Diffusion [16]) are already implemented in MannaSim.

B. The SGT and the MannaSim's Settings

The simulation in NS-2 involves creating scripts in TCL scripting language. This is a tedious and error-prone task, since there are several parameters that need to be adjusted. In order to simplify this task, it was developed an automated system for the generation of TCL scripts used by MannaSim, which is the SGT. Through a friendly interface, the user specifies values for the main parameters of the network. Then the tool takes care of creating the corresponding TCL scripts. The Script Generator Tool is composed by 4 different user interfaces:

Basic Configuration: In this interface we can configure the basic settings of the wireless sensor network simulation like: Transport Protocol (TCP or UDP); Routing Protocol (DSR, TORA, LEACH, Directed Diffusion, DSDV or AODV); MAC (Only IEEE 802.11 is available); Link Layer (Only NS-2 LL default link layer is available); Antenna; Radio Propagation (FreeSpace, Shadowing, ShadowingVis or TwoRayGround); Interface Queue (DropTail, DropTail/XCP, RED, RED/Pushback, RED/RIO, Vq or XCP); Interface Queue Length; Scenario Size; and Simulation Time.

Access Point: Through this interface we can set the AP configurations such as: Number, location, Initial Energy and Transmission Range of the Access Points.

Cluster Head: This interface can be used to configure the Cluster Head settings like: Number, location, Initial Energy and Transmission Range of the Cluster Heads; Transmission Range; Processing Type (Only Aggregate Processing is available); Dissemination Type (Continuous, On Demand or Scheduled); and Dissemination Interval.

Common Node: This interface can be used to set the Common Nodes (CN) settings. Beyond the same parameters that can be configured for Cluster Heads, the Common Nodes can also receive parameters like: Sensing Type (Continuous, On Demand or Scheduled); Sensing Interval; Data Generator Type (Only Temperature and Carbon Monoxide are available, but others can be implemented); Data Average Value; Data Standard Deviation; and Maximum Data Value.

 ISBN: 978-1-61208-398-8

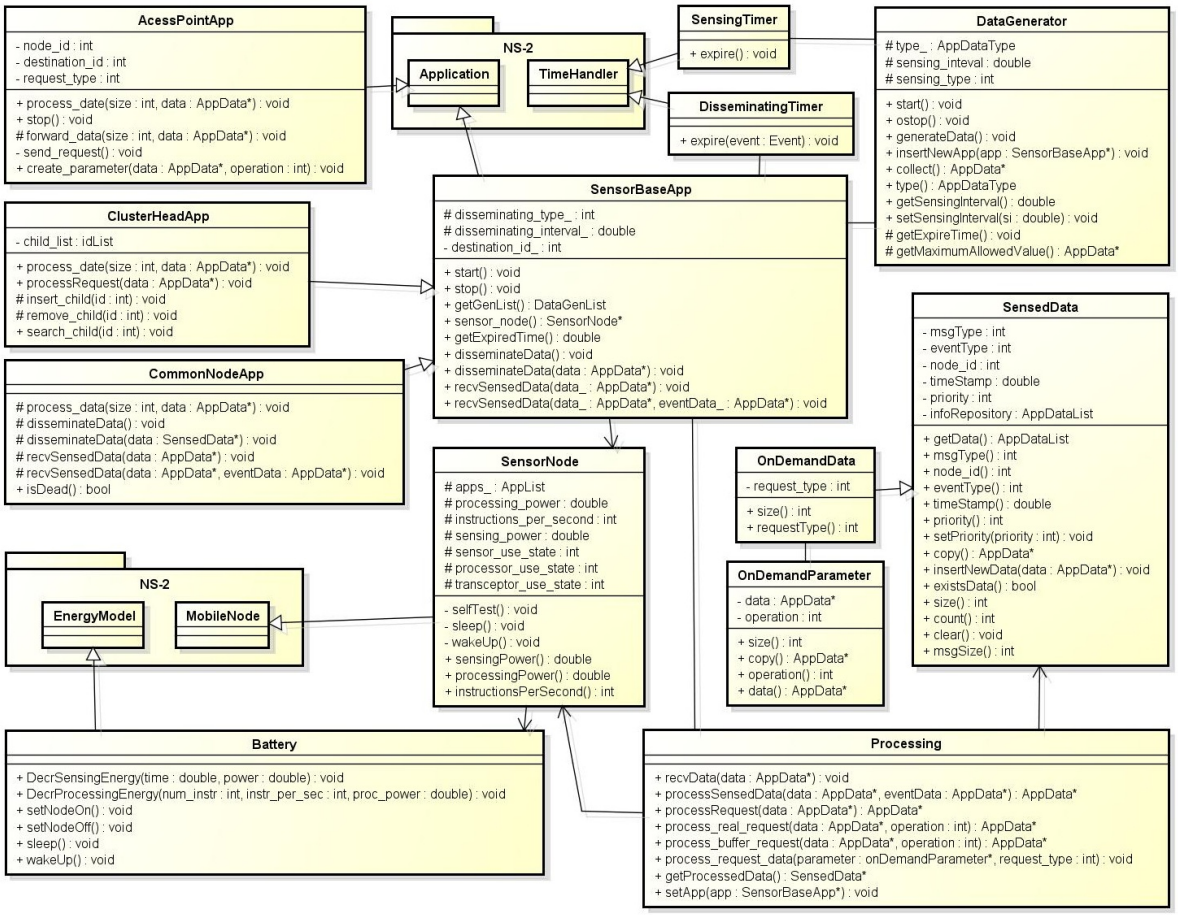


Figure 1. MannaSim's simplified class diagram.

IV. USING MANNASIM FRAMEWORK

To create simulation scenarios using MannaSim, the user must set up the desired parameters of the sensor nodes, and then create node instances and their applications using OTcl language [17]. There are no restrictions regarding the scenarios configuration that may present different compositions, organizations, hierarchical levels, number of nodes, number of access points, and so on.

In order to show how MannaSim works and what kind of results are provided, the following subsections present a particular scenario description and the results provided by MannaSim.

A. The Scenario Description

The particular scenario proposed is a smart home called Follow-Us [2]. This is an application for elderly people monitoring that considers the home instrumentation with WSN technology, clothes manufacturing with wearable computing technology, and the application of concepts involving Social Sensing and Internet of Things. The proposed environment organizes the collected data flow from different types of sensors and networks, establishes a processing routine according to application objectives and provides commands that allow forward the information to be used as control parameters of the environment itself or used by external agents.

Two types of simulation scenarios were developed. The first scenario assumes that the environment is being sensed room by room, and the sensors are on standby while the elderly is in another room. For this scenario, a hierarchical organization in the sensor nodes was used. The second scenario considers that the whole house is being sensed disregarding the rooms divisions, in other words, all the sensors are active all the time. For the second scenario, a flat organization was used in the sensor nodes.

TABLE II. THE SIMULATIONS PARAMETERS

Transport Protocol	TCP
Routing Protocol	AODV
Node Type	MicaZ
Dissemination Type	Scheduled
Dissemination Interval	5 seconds
Initial Energy for Common Nodes	30 Joules
Initial Energy for Access Points	100 Joules
Antenna Range of Common Nodes	10 meters
Antenna Range of Access Point	30 meters
Number of Common Nodes	30
Number of Access Points	3
Scenario Area	30 meters^2

Table II presents the principal parameters used in the simulations. Figure 2 shows how the CNs and the APs were spread over the house in the simulation. Each scenario was executed 33 times, to reach a convergence of the result data, and the simulation time was set to 150 units, time enough to guarantee that the models have been warmed-up sufficiently so that the samples collected will have statistical validity.

 ISBN: 978-1-61208-398-8

Figure 2. The deployment of the sensor nodes and access points over the house scenario.

B. The Results Provided by MannaSim

While the flat organization gives only one group of network for the entire house, the hierarchical organization has to be simulated as eight groups of subnetworks, one for each room of the house that is named in Figure 2. This differentiation is due the fact that in the hierarchical organization each subnetwork is independent and can be inactive for a while. MannaSim provided several results such as described below. It is important to note that only the data shown in this Section was provided by MannaSim's results, the graphs shown in the figures were generated using other tools.

Power Level: Figure 3 shows the power level remaining in the components in each room (for the hierarchical network organization), while Figure 4 presents the power level remaining in the components in each simulation (for the flat network organization).

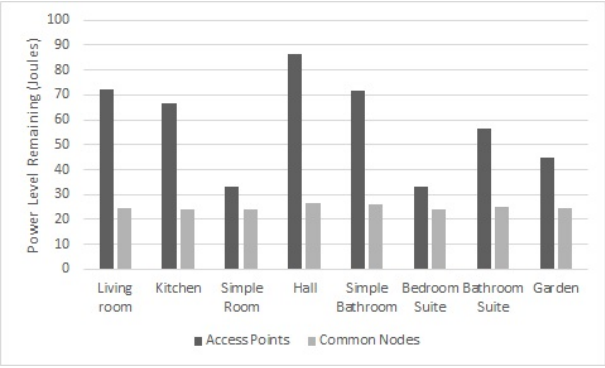


Figure 3. The average power level remaining in the components per room.

Error Types: The average number of errors division in both types of network organization, shown in Table III and Table IV. In this tables the errors are divided by Uninformed, Excessive number of Address Resolution Protocol packets (ARP), The MAC layer was not able to transmit the packet

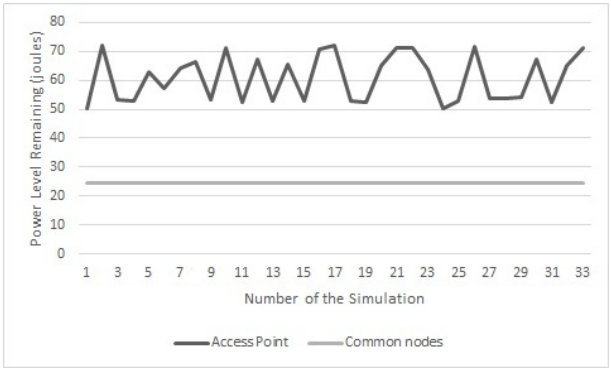


Figure 4. The average power level remaining in the components per simulation.

(CBK), Packet collisions (COL), Excessive packets in the interface queue (IFQ) and Sending packets excessively (RET).

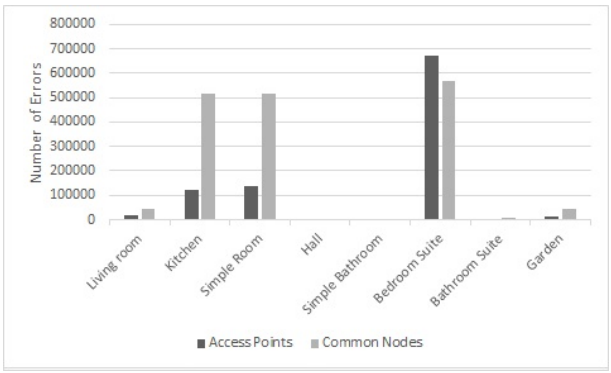


Figure 5. The average errors in the components per room.

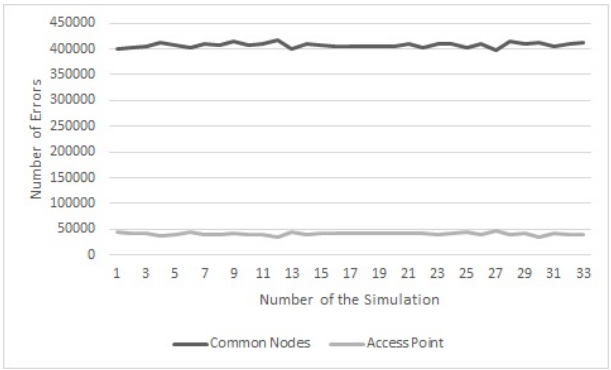


Figure 6. The average errors in the components per simulation.

Number of Errors: The average number of errors in the components in each room (for the hierarchical network organization), presented in Figure 5, or in each simulation (for the flat network organization), presented in Figure 6.

Simulation Events Division: The average events division (Error, Routing, Receiving, Transmitting) in both types of network organization, presented in Figure 7.

Based on the results, it is possible to analyze the WSN from different points of view. In the hierarchical network simulation the average power consumption of the sensor nodes is 5 Joules whereas for the access points is 40 Joules. Although, in the flat network simulation, the average power consumption of the sensor nodes is 5.6 Joules and for the access points is 39

 ISBN: 978-1-61208-398-8

TABLE III. THE AVERAGE ERRORS EVENTS DIVISION IN THE HIERARCHICAL NETWORK ORGANIZATION

Errors Types	Common Nodes	Access Points
Uninformed	891047	0
ARP	11	10
CBK	92	0
COL	63077	1692069
IFQ	6763	0
RET	184	0
Total	961174	1692079

TABLE IV. THE AVERAGE ERRORS EVENTS DIVISION IN THE FLAT NETWORK ORGANIZATION

Errors Types	Common Nodes	Access Points
Uninformed	12619826	21210
ARP	2773	206
CBK	9800	146
COL	603044	1341910
IFQ	193115	0
RET	19000	298
Total	13447558	1363770

Joules. Taking into consideration these results, it is possible to conclude that the power consumption is almost the same in both scenarios.

The average number of errors in the flat network simulation is almost 6 times superior that the average number of errors events in the hierarchical network. Whereas the hierarchical network got a total of 2.653.253 errors, the flat network got an average of 14.811.328. In Table III, for the hierarchical network organization, we can see that packet collisions are the most frequent errors, followed by uninformed errors. Meanwhile, Table IV shows that, for the flat network organization, most part of the errors are uninformed and, after that, packet collisions are the most frequent errors.

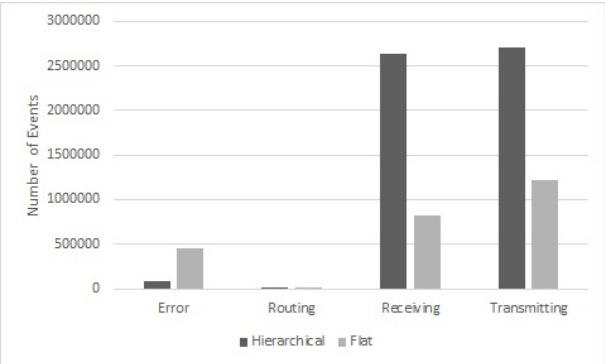


Figure 7. The average events division in the simulations

Figure 7 indicates that if we compare in proportion terms, error events were only 1% of the total events in the hierarchical network, while in the flat organization network, errors took approximately 18% of events. This information shows that, giving the chosen configurations, the hierarchical network would bring a smaller error percentace than flat network, therefore, it would be more appropriate to use in the simulated application proposed.

The impacts of different network organizations over the total errors is a kind of information that cannot be observed by other simulation tools, such as the ones presented in related works, even SensorSim, which also extends NS-2 core. The principal disadvantage of MannaSim is that the power consumption module is under development, so it does not considers factors such as the executed code in the sensor nodes to make its estimatives yet.

V. CONCLUSIONS AND FUTURE WORKS

Select the correct level of detail (or level of abstraction) for a simulation is a difficult task. Few details can produce simulations that are misleading or incorrect. On the other hand, add to many details requires more time to implementation and debugging, which probably will slow down the simulation and distract from the research problem. In this meaning, it is relevant that there is a simulation tool that helps the WSNs developers to build their scenarios, as well as, performs their experiments in order to obtain results that permits proves their theories. The MannaSim framework allows that different WSN scenarios can be simulated, offering possibilities for the configuration of WSNs and applications. Furthermore, it provides a set of base classes that can be extended, making the framework specific to the needs of researchers who use it.

The NS-3 tool, a later version of NS-2, whose core is used by MannaSim, was published in [18]. The migration of MannaSim's core to NS-3 is a future work point. However, as the NS-3 project was started "from the beginning", this migration is subject to a study of the changes needed in MannaSim's set of classes so that the interface with NS-3 can be made. Furthermore, the migration should not be made while there are no scientific studies that prove the efficiency and effectiveness of the NS-3.

As future works in MannaSim, it can also be cited the construction of a realistic battery energy decay model, considering the influence of ambience temperature in the battery capacity. To improve the power of observation of the simulation results, we intend to create a graphical interface for the MannaSim's output. Beyond this, other features, like comunication interference between the nodes, will be implemented. Finally, a more ambitious project intends to add to MannaSim the feature of Instruction Level simulation.

ACKNOWLEDGMENT

We would like to thank Thais Regina de Moura Braga and Fabricio Aguiar Silva for their commitment to the development of MannaSim Framework. We also thank the National Council for Scientific and Technological Development (CNPq) from the brazilian government for it is financial support on the project (Project process number 55.2111/2002-3).

REFERENCES

[1] V. Gungora, B. Lu, and G. Hancke, "Opportunities and challenges of wireless sensor networks in smart grid," IEEE Transactions on Industrial Electronics, vol. 57, no. 10, October 2010, pp. 3557–3564.

[2] M. L. A. Ghizoni, A. Santos, and L. B. Ruiz, "Follow-us: A distributed ubiquitous healthcare system simulated by mannasim," Computational Science and Its Applications, vol. 7336, June 2012, pp. 588–601.

[3] T. Torfs, T. Sterken, S. Brebels, and J. Santana, "Low power wireless sensor network for building monitoring," IEEE Sensors Journal, vol. 13, no. 3, March 2013, pp. 909–915.

[4] R. A. Khan, S. A. Shah, and M. A. Aleem, "Wireless sensor networks: A solution for smart transportation," Journal of Emerging Trends in Computing and Information Sciences, vol. 3, no. 4, April 2012, pp. 566–571.

 ISBN: 978-1-61208-398-8

[5] G. N. L. R. Teja1, V. K. R. Harish, N. M. Khan, R. B. Krishna, R. Singh, and S. Chaudhary, "Land slide detection and monitoring system using wireless sensor networks (wsn)," in IEEE International Advance Computing Conference, Gurgaon, February 2014.

[6] C. K. Ng, C. H. Wu, L. Wang, W. H. Ip, and J. Zhang, "An rfid-enabled wireless sensor network (wsn) monitoring system for biological and pharmaceutical products," in International Symposium on Computer, Consumer and Control, Taichung, June 2014.

[7] "The network simulator (ns-2) home page," http://www.isi.edu/nsnam/ns/, accessed on November, 2014.

[8] L. B. Ruiz, J. M. Nogueira, and A. F. Loureiro, "Manna: A management architecture for wireless sensor networks," IEEE Communications Magazine, vol. 41, no. 2, February 2003, pp. 116–125.

[9] G. Chen, J. Branch, M. Pflug, L. Zhu, and B. Szymanski, "Sense: A wireless sensor network simulator," in Advances in Pervasive Computing and Networking, 2005, pp. 249–267.

[10] S. Park, A. Savvides, and M. B. Srivastava, "Simulating networks of wireless sensors," in Proceedings of the Winter Simulation Conference, vol. 2, Arlington, December 2001, pp. 1330–1338.

[11] B. Titzer, D. Lee, and J. Palsberg, "Avrora: Scalable sensor network simulation with precise timin," in Proceedings of the 4th International Symposium on Information Processing in Sensor Networks, April 2005, pp. 477–482.

[12] A. Boulis, "Castalia: Revealing pitfalls in designing distributed algorithms in wsn," in Proceedings of the 5th International Conference on Embedded Networked Sensor Systems, 2007, pp. 407–408.

[13] L. Shu, M. Hauswirth, H.-C. Chao, M. Chen, and Y. Zhang, "Nettopo: A framework of simulation and visualization for wireless sensor networks," Ad Hoc Networks, September 2010, pp. 799–820.

[14] "The mannasim's home page," http://www.mannasim.dcc.ufmg.br/, accessed on December, 2014.

[15] B. Wang, C. Shen, and J. Li, "Study and improvement on leach protocol in wsns," Automatic Control and Artificial Intelligence, March 2012, pp. 1941–1943.

[16] N. El-Bendary, O. Soliman, N. Ghali, A. E. Hassanien, V. Palade, and H. Liu, "A secure directed diffusion routing protocol for wireless sensor networks," Next Generation Information Technology, June 2011, pp. 149–152.

[17] "Otcl home page," http://otcl-tclcl.sourceforge.net/otcl/, accessed on April, 2015.

[18] T. R. Henderson, M. Lacage, and G. F. Riley, "Network simulations with the ns-3 simulator," in Proceedings of the Special Interest Group on Data Communication Conference, Seatle, Washington, August 2008.

 ISBN: 978-1-61208-398-8

Comparative Analysis of the Algorithms for Pathfinding in GPS Systems

Dustin Ostrowski

Metegrity Inc.
Edmonton, Canada
e-mail: dustin.ostrowski@gmail.com

Iwona Pozniak-Koszalka, Leszek Koszalka, and Andrzej Kasprzak
Wroclaw University of Technology
Wroclaw, Poland
e-mail: {iwona.pozniak-koszalka, leszek.koszalka, andrzej.kasprzak}@pwr.edu.pl

***Abstract*—The objective of the paper was to determine which search method is suitable for implementation in GPS systems. The properties of pathfinding algorithms were tested and discussed taking into account this type of systems. Six algorithms have been evaluated, including three different implementations of Dijkstra algorithm, Bellman-Ford algorithm, A* star algorithm, and bidirectional Dijkstra's algorithm. Simulation experiments were carried out using the real digital maps and with the designed and implemented experimentation system. Studies were performed with respect to various parameters. After thorough examination and interpretation of conclusions, the algorithms which fit to GPS systems were selected.**

***Keywords*—GPS; search algorithms; experimentation system; path finding**

I. INTRODUCTION AND MOTIVATION

Nowadays we are living in the age where new technologies, innovations, great inventions such as the location with GPS have become an integral part of everyone's life. It is hard to imagine how the world could function without the GPS systems. The number of the GPS users is increasing very rapidly [1]. One of the major advantages of the GPS over the traditional searching the route to the target destination is the speed of action. GPS system consists of two basic components – digital maps and the shortest path search algorithm [2] and [3]. The ordinary user does not even realize that algorithmics surrounds him.

There are proposed in literature many search algorithms for solving pathfinding problem e.g., [1][3[4]. In some works, the properties of these algorithms are evaluated [4]. In this paper, the six search algorithms implemented by the authors aretested and evaluated. All the research was carried out on real digital maps. Based on the implemented experimentation system (programs in C++) and properly designed scenarios – the authors made a comparison of the results produced by these algorithms. The main objective of this paper was to find an algorithm that is most suitable for GPS systems.

The paper is organized as follows. In Section II, we formulate the formal model of the considered problem. Section III contains the description of the considered pathfinding algorithms. The experimentation system is presented in Section IV. The results of the simulation experiments are discussed in Section V. The final remarks, conclusion, and suggestions to further research in the area appear in Section VI.

II. PROBLEM STATEMENT

A. Mathematical model

The shortest path problem is an issue consisting in finding the shortest connection between vertices in the weighted graph [2]. There is a directed graph G (V, E), whereV denotes the set of vertices, E is the set of edges connecting vertices.

The weighted function $w: E \to R$ is assigning real-valued weights to the edges – the weights canbe interpreted as costs to traverse the edges. The total cost of a path $p = (v_0, v_1 \dots v_k)$ is defined by (1):

$$c(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i) \tag{1}$$

The cost of the shortest path from u to$v$ can be expressed in the form given by (2):

$$\delta(u,v) = \begin{cases} \min\{c(p): u \xrightarrow{p} v, & \text{if path from } u \text{ to } v \text{ exists} \\ \infty, & \text{otherwise} \end{cases} \tag{2}$$

The shortest path from u to$v$ is each path p *from* u to v fulfilling the condition defined by (3):

$$c(p) = \delta(u,v) \tag{3}$$

More information about the shortest path problem is widely available, e.g., in [3] and [4].

B. Representation of the graph in a computer system

In the considered problem, the map can be represented as a huge graph [5].In such a model, the edges can be represented by means of roads and the vertices can be represented by intersections. In this work, a graph is represented by a list of incidence L. In the *n-th* list are stored incident vertices with the *m-th* vertex. Each element

 ISBN: 978-1-61208-398-8

of the list contains the ID of vertex, which is connected and a weight of a proper edge. The list has only as many elements as the number of possible connections for a given vertex what allows saving a lot of memory resources. This is the recommended solution for GPS system – it significantly increases the efficiency and speed of operations.

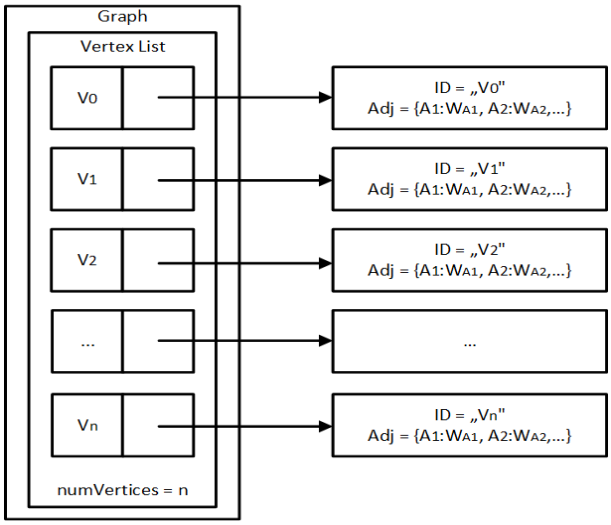


Figure 1. Representation of incidence list L.

This type of list is presented in Figure 1, where:

n – Number of vertices in graph,
ID – Given single vertex in graph,
Adj – Adjacency vertices to a given ID,
W – Weight of given adjacency vertex.

C. The weighted functions

The most commonly used weighted functions in GPS system correspond to the shortest route case and the fastest route case [6].

In the shortest route case, the weighted function is defined as the sum of the distances between successive intersections on the route. The total cost can be described by (1) in which $w(v_{i-1}, v_i) = d(v_{i-1}, v_i)$, where *d (u, v)* is the distance between the intersection *u* and *v*, i.e., by (4):

$$c(p) = \sum_{i=1}^{k} d(v_{i-1}, v_i) \quad (4)$$

In the fastest route case, the weighted function is defined as the sum of ratios of the distances between successive intersections and speed limits on the road between intersections. The total cost can be described by (5), where *s (u, v)* describes speed limit between *u* and *v*.

$$c(p) = \sum_{i=1}^{k} \frac{d(v_{i-1}, v_i)}{s(v_{i-1}, v_i)} \quad (5)$$

Search assumption is that that we are looking for the shortest or fastest route from the starting point (source) to the destination point – it is only one destination point. An additional assumption is that we are dealing with a map, so the edges in the graph do not have negative weights.

III. ALGORITHMS

A. Dijkstra's algorithm

Dijkstra's algorithm is used for solving a single source shortest path problem. It can find the path with the lowest cost between an initial vertex and every vertex in the graph. It can also find the shortest path from single vertex to single destination point by stopping the algorithm when the final shortest path was found – this way was applied in the designed experimentation system (simulator) allowing comparing this algorithm with the other considered algorithms. In our case, the graph should be directed, weighted and the edges should have non-negative weights. The basic idea of the algorithm lies in the fact that the information about predecessors is stored together with the information about the shortest path to a given vertex. In our implementation, we maintain a priority queue of vertices that provides three operations [6]and [7]:

Step 1. Inserting new vertices to the queue.
Step 2. Removing the vertex with the smallest distance.
Step 3. Decreasing the distance value of some vertex during relaxation.

Priority queue affects the performance of the algorithm in very large extent. For this reason, three different ways of implementing a queue were used what allow creating three versions of the algorithm.These versions are:

1) Priority queue as an array.
2) Priority queue as a binary heap.
3) Priority queue in the form of the Fibonacci heap [8].

B. Bellman-Ford

The Bellman-Ford algorithm solves the shortest path problem basing on the relaxation. The algorithm iteratively generates a better solution from a previous one until it reaches the best solution. It activates two loops, one running n-1 iterations and the other going through all edges. Bellman-Ford is slower than Dijkstra's in most cases but it can be more useful in some cases [3]. The details about this algorithm can be found, e.g., in [2] and [7].

C. Bidirectional Dijkstra's algorithm

This algorithm searches simultaneously from the source vertex (node) onward and from the destination vertex (node) backwards and stop when the two routes meet in the middle [9]. Searching the shortest path can be divided into two stages. When two paths meet at one vertex it is advisable to check whether the current path is the shortest. The principle consists of saving weight s of thefounded paths(in the

 ISBN: 978-1-61208-398-8

tables). If the weight is less than the sum of the values in the tables (for both instance of the algorithm) for vertices at the beginning of both priority queues – then the path is the shortest. Binary heap was used in our implementation as a priority queue. Searching from both the source and destination in a homogenous graph can reduce the search space to approximately half the size compared to only searching from the source.

D. A algorithm*

A star (A*) algorithm is a heuristic algorithm [7]. The algorithm allows findingan approximated solution - the least cost path from the start point to the destination point. A star uses a distance-plus-heuristic function*f(x)* to determine the order of visiting the nodes in a tree. This function can be defined by (6):

$$f(x) = g(x) + h'(x) \quad (6)$$

where *g(x)* – is the total distance it has to be taken to get the current position *x*, *h'(x)* – is the estimated distance from the current position *x* to the destination point. Such a heuristic is used to estimate on how far away it will take to reach the destination.

IV. EXPERIMENTATION SYSTEM

The experimentation system was created in order to properly investigate the properties of the pathfinding algorithms. It contains the programmed simulator with thesix implemented algorithms. It also ensures creation of the experimental scenarios – maps. Twoapplications (modules of the system) are available:

(I) *Real maps module*– allowing for creating a graph based on the loaded map, what gives possibilities of using the actual digital maps in order to reproduce the real conditions.

(II) *Arbitrary maps module*– allowing for generating an arbitrary created graph - with the chosen number of vertices, density, start point, number of iteration, etc. This part can be used to make simulation experiments in automatic way.

The block-diagrams of the functionality of these modules ofthe experimentation system are presented in Figure 2 (module I) and Figure 3 (module II).

Input problem parameters of the module I are:

U1 – Starting point as the node (*U*) marked on the map area.

U2 – Destination (End) point as the node (*V*) marked on the map area.

U3 – List of incidence *L* (see an example in Figure 1).

U4 – Weighted functions(weights assigned to the edges).

Output parameters of the moduleIare:

Q1 – Path found - denoted by $\delta(u, v)$- see formula expressed by (3).

Q2 – The total cost (length/time) of the founded route.

Q3 – The average execution time of the algorithm.

In both modules, the algorithm is treated as a special input. The experimentation system gives opportunities to use the following six algorithms:

- A star (A*),
- Dijkstra,
- Dijkstra binary,
- Dijkstra bi-directional,
- Dijkstra Fibonacci,
- Bellman-Ford.

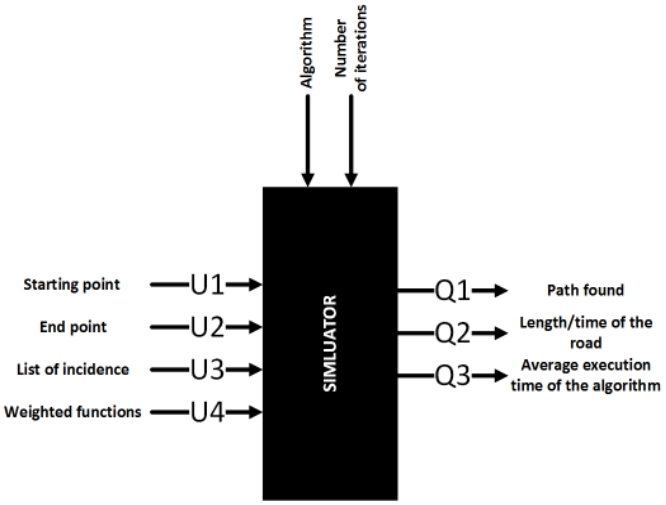


Figure 2. Input/output –module I - with real maps.

Input parameters of the module II are:

U1 – Starting point as the node (*U*) marked on the map area.

U2 – Destination (End) point as the node (*V*) marked on the map area.

U3 – Density of graph. This value expresses the amount of edges to remove (in relation to the full graph).

U4 – Hash - the value used for the pseudorandom number generator.

U5 – Weighted functions (weights assigned to the edges).

U6 – List of incidence *L* (see an example in Figure 1).

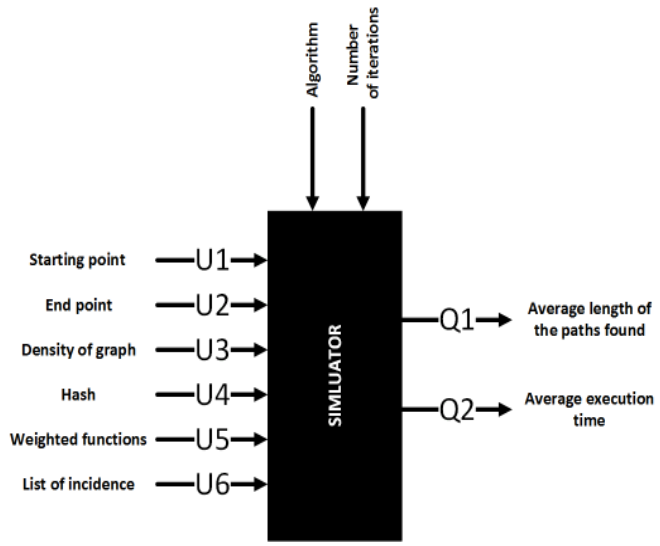


Figure 3. Input/output – module II – arbitrary maps.

Output parameters of the moduleII are:

- Q1 – The total cost (length/time) of the founded route.
- Q2 – The average execution time of the algorithm.

To investigate and compare the algorithms OsmGPS application was used. The tool has been implemented in C# environment. The main application window for the module I can be seen in Figure 4.

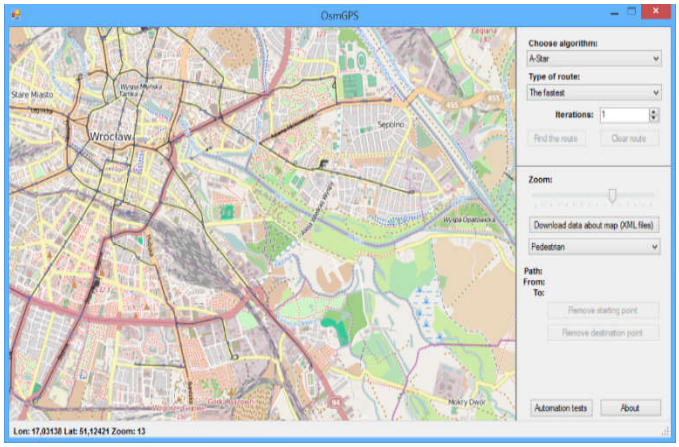


Figure 4. Main window - real map.

The main application window for the module II is shown in Figure 5.

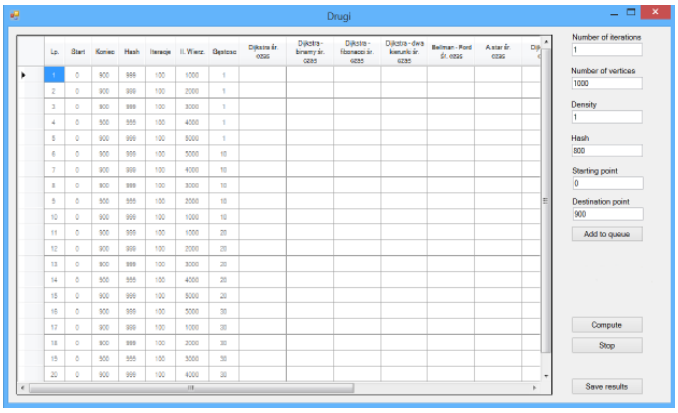


Figure 5. Main window–arbitrary map.

Both modules allow searching for the shortest and fastest routes. The experiments were made on MS Windows 8.1.

V. INVESTIGATION

A. *Experiment # 1*

The aim of Experiment #1 was checking the relationship between the execution time and the number of vertices in the graph. The locations of the starting points and destination points were always the same. Only the area of the map was changed(by increasing the number of vertices and edges).The results are presented in Figure 6 and Figure7.

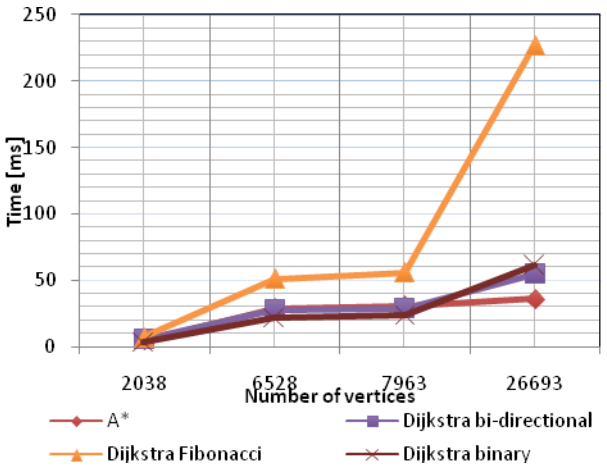


Figure 6. Execution time and number of vertices – the fastest algorithms.

 ISBN: 978-1-61208-398-8

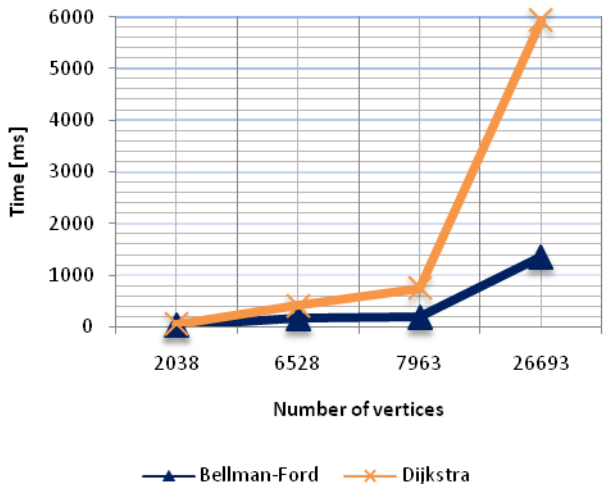


Figure 7. Execution time and number of vertices – the slowest algorithms.

As it can be seen in Figure 6, the A-star reached the lowest execution times. Dijkstra binary priority queue and Dijkstra bi-directional both reached a little bit worse times. The mentioned algorithms get the best times for graphs with high and low number of vertices. Two remaining algorithms (Bellman-Fordand Dijkstra) achieved much worse results.As expected, the execution timeswere raising a lot with increasing amounts of vertices in the graph. All the algorithms get similar execution time for graphs with less than the number of 6528 vertices.

B. *Experiment # 2*

The objective of this experiment was checking the relationship between the execution time and the path length. In the same area (constant number of vertices and edges) was set starting point – always the same (constant). The destination point was systematically moved away. The results are presented in Figure 8 and Figure9.

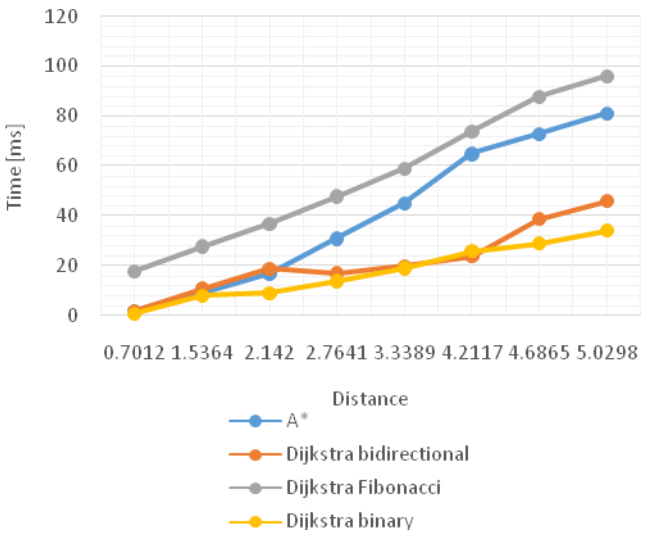


Figure 8. Execution time and path length – the fastest algorithms.

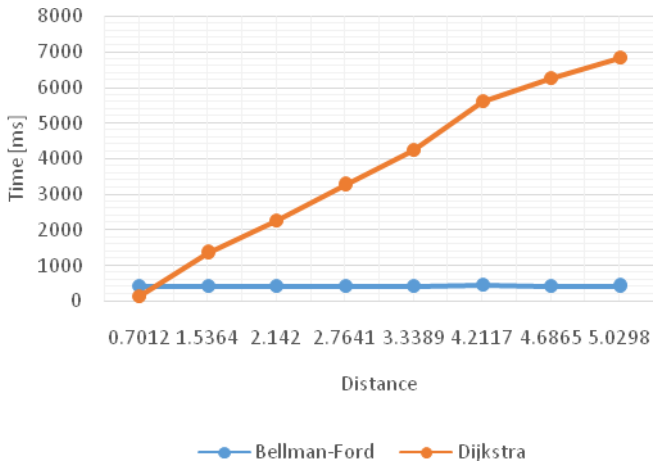


Figure 9.Execution time and path length – the slowest algorithms.

It can be seen in the figures that the execution time is growing steadily with the increasing distance. The Dijkstra based on binary queue and Dijkstrabi-directional were the fastest. For short distances, time differenceswere imperceptible. Only Dijkstra Fibonacci achieved worse results at each stage of the test. A-star and Dijkstra algorithm were apparently slower in comparison to others.

Bellman-Ford and Dijkstra again were the slowest of all. The execution time of Bellman-Ford algorithm was almost the same for different distances because of constant number of checks. This is due to the fact that algorithm searches paths to each of vertices and is not aborted before. The execution time for Dijkstra algorithm was constantly growing with increasing distance.

C. *Experiment # 3*

This experiment was a similar test to the previous experiment. It was performed for the area of Wroclaw city but this time the starting point and the end point were gradually moved close towards each other. Initially starting point was set in the southern part of the city and end point in the northern part of the city. In the middle of the founded route was the center of the city – a large collection of edges and vertices. The results are presented in Figure 10 (logarithmic scale).

 ISBN: 978-1-61208-398-8

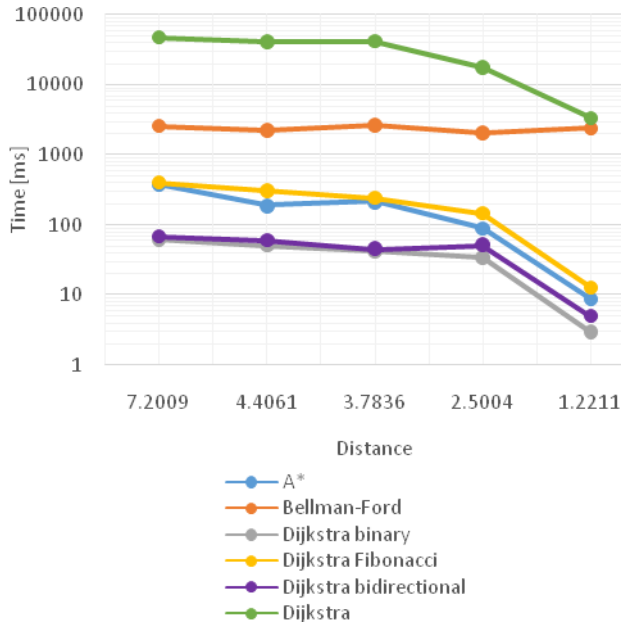


Figure 10. Relationship between execution time and distance – points are moved close towards each other.

It can be observed in Figure 10, that the best results have been produced by Dijkstra based on binary queue, Dijkstrabidirectional and A* algorithm. Very good performance of Dijkstra bidirectional should not be surprising in this test. This is due to the principle of the algorithm where it searches simultaneously from starting point onward and next from destination point backwards, and it stops when the two paths meet in the middle. Approaching the starting point and the destination point we facilitate the work of the algorithm. Once again the worst algorithms were: Bellman-Ford and simple Dijkstra.

D. Experiment # 4

The objective was to verify the performance of algorithms based on the graphs with given amounts of vertices, destination point, starting point, the density and the number of iterations. In this study, the attention was focused on the density of graphs. Tests were performed for different number of vertices but with the same conditions, e.g., for the same graph. The results are presented in Figure 11.

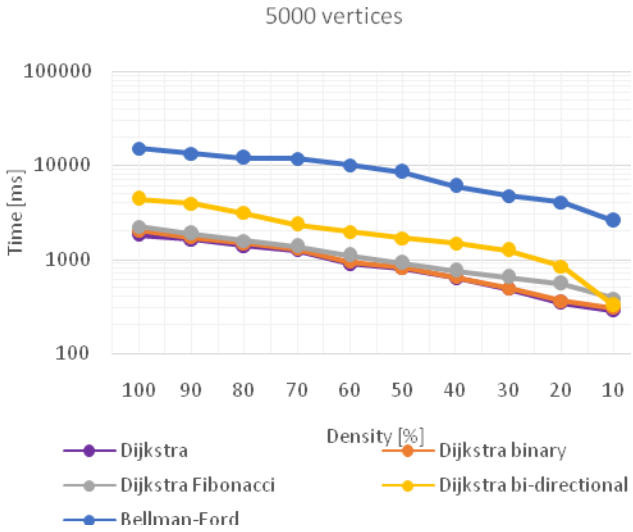


Figure 11. Relationship between density of graph and execution time - graph with 5000 vertices.

E. Experiment # 5

The aim of this experiment was toshow the relations between execution time and the number of vertices for 100% density.The results are presented in Figure 12.

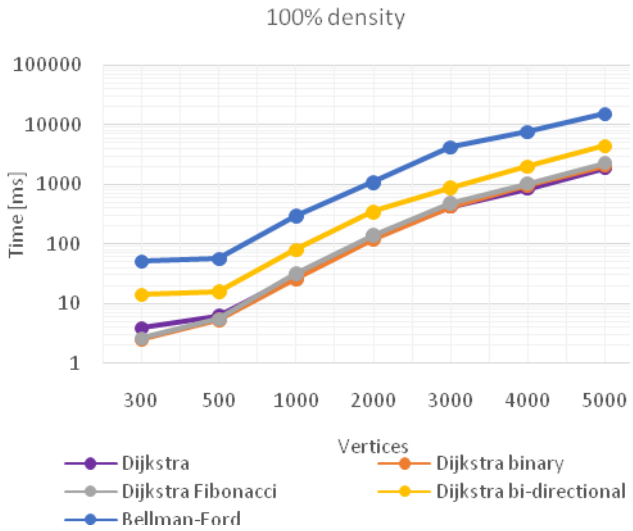


Figure 12. Relationship between vertices and execution time – density 100%.

Figure 12 demonstrates that the Dijkstra's binary achieved the best results for graphs with high density. Competitive for this algorithm were Dijkstra Fibonacci heap and simple Dijkstra algorithm.

VI. FINAL REMARKS AND PERSPECTIVES

Based on the simulation experiments, we may say that for GPS system the Dijkstra algorithm with priority queue in the form of the binary heap performed as the best. This algorithm achieved very good results almost in each

experiment – no matter how big the graph was. Also, easy implementation is a big strength of this method. However, Fibonacci heap definitely did not work well in graphs with high density. This is due to the structure of the heap (many nodes with pointers to the neighbors, parents, list of children). In binary heap we have only the array of elements. It is worth to mention that the bidirectional Dijkstra algorithm also achieved good results in many cases, in particular for graphs with low density. Using this algorithm can be recommended as a good solution in GPS systems because the maps have usually density at approximately a few percent levels.

The authors are planning to focus on the algorithms based partially on the evolutionary approaches, e.g., presented in [10] and [11]. There are also several interesting issues that might be considered in the future work, includingmore complex experiments with: more exact/heuristic algorithms, larger topologies, and detailed analysis of computational complexity of algorithms, as well as memory usage, following the ideas of multistage experiment design presented in [12].

ACKNOWLEDGEMENT

This work was supported by the statutory funds S40029_ K0402, Wroclaw University of Technology, Poland.

REFERENCES

[1] Royal Pingdom, “Google Maps turns 7 years old – amazing facts and figures” [retrieved: July, 2014] at :http://royal.pingdom.com.

[2] A. Kasprzak, Wide Area Networks, OWPW Publishing House, Wroclaw, 2001 /in Polish/.

[3] J. Larsen and J. Clausen, “The shortest path problem” [retrieved: February, 2015] at: http://imada.sdu.dk/~jbj/DM85/lec6a.pdf

[4] D. Johansson, “An evaluation of shortest path algorithms on real Metropolitan Area Network”, Linköpings Universitet, Report: SE-581 83, Linköping, Sweden, 2008.

[5] J. Koszalew, “Data structure for graph representation: selected algorithms,”[retrieved: July, 2014] at: http://asdpb.republika.pl.

[6] W. Lung and D. Tseng, “Graph theory: shortest path”, [retrieved: June, 2014] at: http://www.cs.cornell.edu/~wdtseng/icpc/notes.pdf.

[7] T. H. Cormen, Algorithms Unlocked, MIT Press, Cambridge, 2013

[8] K. Wayne, “Fibonacci heaps” [retrieved: January, 2015] at: http://cs.princeton.edu /~wayne/cs423/.

[9] G. Vaira and O. Kurasova, “Parallel bidirectional Dijkstra’s shorthest path algorithm”, Frontiers in Artificial Intelligence and Applications, vol. 224, 2011, pp. 422-435.

[10] T. Miksa, L.Koszalka, and A. Kasprzak,“Comparison of heuristic methods applied to optimization of computer networks”,Proc. of 11th Intern.Conf. on Networks, ICN 2012, IARIA, pp. 34-38.

[11] D. Ohia, L. Koszalka,and A. Kasprzak, “Evolutionary algorithm for solving congestion problem in computer network”, LNCS, Springer, vol. 5711, 2009, pp. 112-121.

[12] A.Kakol, I.Pozniak-Koszalka, L.Koszalka, A. Kasprzak, and K.J. Burnham, “An experimentation system for testing bee behavior based algorithm for a transportation problem”, LNCS, Springer, vol. 6592, 2011, pp.11-20.

 ISBN: 978-1-61208-398-8

A Study on GPS Positioning Method with Assistance of a Distance Sensor

Yasuhiro Ikeda, Hiroyuki Hatano, Masahiro Fujii, Atsushi Ito, Yu Watanabe
Graduate School of Engineering,
Utsunomiya University
e-mail:{ikeda@degas., hatano@, fujii@, at.ito@, yu@}is.utsunomiya-u.ac.jp

Tomoya Kitani
Graduate School of Informatics,
Shizuoka University
e-mail:t-kitani@ieee.org

Toru Aoki
Research Institute of Electronics,
Shizuoka University
e-mail:rtaoki@rie.shizuoka.ac.jp

Hironobu Onishi
Graduate School of Engineering,
Shizuoka University
e-mail:onishi@hatanolab.eng.shizuoka.ac.jp

***Abstract*—In order to estimate a receiver's position, a Global Positioning System (GPS) receiver has to receive the signals from at least four observable satellites. However, in urban areas, the number of the observable satellites decreases because urban areas have many buildings. In such cases, the position estimation cannot be performed well. So, our research considers position estimation in case that the observable satellites are decreased. Many vehicles have the distances sensors which can measure traveling distance. Our proposal can estimate own position by using the traveling distance and the previous position which is estimated under good GPS condition. Our experimental results will show our effectiveness of the proposal.**

Keywords*–GPS; Positioning; Urban canyon; Lack of observable satellites.

I. Introduction

In recent years, location-based services have been increasing. Many of these services require the information of user's position. Car navigation systems can be given as an example. In general, these services use the Global Positioning System (GPS). GPS is a system that can estimate user's position by using flying satellites around the Earth.

The position estimation by GPS calculates the position, based on the measurement of the distances between the GPS receiver and the GPS satellites. In order to perform the position estimation, GPS receiver requires at least four satellites which can be received in line-of-sight [1]. However, in urban areas, the number of the observable satellites decreases because of the buildings. These situations are called urban canyon. In urban canyon, GPS signals from observable satellites tend to degrade because of multi-path propagation. The signals via multi-path should not be used because the errors are included in the propagated paths. In order to avoid multi-path signals, the number of direct-path satellites which we prefer to use is more decreased. In such cases, the position estimation cannot be performed well. For example, the estimator cannot estimate the receiver's position if the number of observable satellites becomes less than three. Or, the performance becomes worse compared to the conditions in open sky where we can observe more than four satellites. Actually, conventional GPS systems rely on other information such as map information, base stations of 3G cellular networks, or Wi-Fi networks against the bad GPS measurement. The positioning method which does not rely on other information is desired because the simple positioning should not use big data and wireless links to get other information.

So, our research considers the position estimation in the case that the observable satellites are decreased. Many vehicles have wheel rotation sensors which can measure the traveling distances. We propose the novel positioning method which uses the information of the previous traveling distance [2][3]. Our proposed method uses the information of both the distance sensor and the previous position which is estimated under good condition, to GPS positioning. Higher accuracy in position estimation is expected by our proposed method, even when the number of observable satellites is decreased. Also, it can be expected to prevent the position estimation impossibility when the number of the observable satellites becomes three.

In this paper, we will show the case that the number of the observable satellites becomes three as the worst case to introduce and evaluate our proposed method. Here, the position estimation is performed by traveling on a bicycle. Using experimental results, we will show the effectiveness of our proposed method.

This paper organized as follows. In Section II, we will introduce related works briefly. In Section III, we will show our proposal in detail. In Section IV, we will present the proposal's performance by field experiments. Finally, Section V summarizes the paper.

II. Related Work

In this section, we will show the related methods to improve the accuracy of the position estimation. Here, we are introducing the traditional and basic technologies.

A. Differential GPS (DGPS)

The Differential GPS (DGPS) is the method for improving the position estimation accuracy [4]. The DGPS uses the GPS base station. The GPS base station transmits the information of the error amount in the GPS measurement to near GPS receivers. Measuring of the error information is performed accurately at the GPS base station.

 ISBN: 978-1-61208-398-8

Generally, the position estimation by GPS calculates the position by using the measurement of the distances between the GPS receiver and the GPS satellites. However, some errors are included in the distance measurement. The distance errors by clock difference, the ionospheric delay, and the troposphere delay can be given as examples. The estimation accuracy of user's position can be improved by correcting the error information which is generated at the GPS base station. However, the GPS estimation is used in various locations, such as urban areas, rural areas, sea, and mountains. In the urban areas, the DGPS cannot correct the propagation delay caused by the reflection at buildings. Therefore, in such case, the position estimation cannot be performed well.

B. Dead Reckoning (DR)

The Dead Reckoning (DR) is the method of performing position estimation by the information of the relative movement [5]. In other words, DR uses the information how much we traveled from the previous position. Since the DR does not require any infrastructures, the DR is not limited to any area.

In vehicles, various sensors exist to detect the direction. Usually, the angular velocity is detected by the angular velocity sensor. The angular velocity can be calculated by integrating the traveling direction [6]. Also, it is possible to detect the direction by using a fiber optic gyroscope. The fiber optic gyroscope is a device for determining the direction by measuring the time difference of the light when the angular velocity is added to the fiber optic.

By using vehicle speed pulses, it is possible to detect the traveling distance. The vehicle speed pulse is the signal which is generated according to the rotational speed of the drive shaft of the vehicle. We can measure the traveling distance based on the circumference of the tire and the vehicle speed pulse.

In the DR, the relative position can be estimated by using the moved direction and the traveled distance. The DR is often used with a map matching technique, as described in the next section. The combination of both DR and map matching are often used in the car navigation systems.

C. Map Matching

The map matching is the method for finding the appropriate position of the vehicle on the road by using the map information [7]. It is used in combination with the position estimation methods such as GPS and DR.

Currently, the map matching and the DR are commonly used in the car navigation systems [8]. The DR is a system for determining the relative position from the previous position. In DR, the error accumulates when the error occurs in the distance sensor and the direction sensor. This problem can be solved by using the map matching, but the map matching cannot be used in a place which does not have the map information. So, in order to estimate the absolute position, the DR with the map matching is often used with GPS. However, the sensor data from the DR does not improve the positioning accuracy of GPS directly.

As a new method, we want to improve the position estimation accuracy by adding the sensor data, to the GPS position estimation directly. Here, we use a distance sensor. We proposed the positioning method which can estimate the absolute position by the combination of the distance sensor and GPS [2][3]. By our proposed method, the absolute position can be estimated even if only GPS cannot estimate own position because of the bad environment. The bad environment for the conventional GPS is under lack of the observable satellites. This case often happens in urban areas. In our proposal method, we can keep estimating the user's absolute position by assistance of the distance sensor.

III. Proposed Method

In this section, first, we will show the problems of the position estimation by GPS. Thereafter, we will show our proposed method which uses the distance sensor and the previous position information.

A. Problems of Position Estimation by GPS

The GPS positioning estimates the receiver's position based on the distances between the receiver and the satellites [9]. For estimating the 3D position of the receiver, the receiver needs three relations, that is, three satellites. Moreover, the receiver needs to estimate own clock error because the general receivers are equipped with inexpensive crystal clocks. Therefore, the GPS position estimation needs four relations, that is, four satellites. For carrying out the positioning calculation, the GPS position estimation needs at least four observable satellites. The observable satellite means the satellite which can receive its signal in line-of-sight. However, in urban areas, the number of the observable satellites is decreased because the receiver cannot observe the low elevation satellites due to the interference of buildings. Therefore, the receiver is not always possible to observe more than or equal to the four satellites in line-of-sight. So, the number of the observable satellites tends to decrease. Such decrement results in degradation of the position estimation accuracy. As a worst case, the receiver cannot estimate its own position if the number of the observable satellites becomes less than four. This is the most common problem in urban area.

To solve the above problem, we propose the novel GPS estimation which uses the distance sensor and the previous position information. The proposal method uses the previous receiver's position. We can measure the traveling distance from the previous position by the distance sensor. We assume the previous position as the quasi-satellite which uses both the previous position and the traveled distance. By the proposed method, we expect that the position estimation is possible even if the number of the observable satellites are three. In addition, we also expect the improvement of position estimation accuracy when the number of the observable satellites are low. The detailed procedure is shown in the next subsection.

B. Proposed Position Estimation Algorithm

The proposed method is assumed to be used in position estimation on vehicles such as cars or bikes. The assumed situation of our research is shown in Figure 1. We consider 2 observation times (the time t and $t + \tau$). The position coordinate of the receiver at the time t is (x_0, y_0, z_0). And the position coordinate of the receiver at the time $t + \tau$ is (x, y, z). We want to estimate the position (x, y, z). We define that the position of the i-th satellite is (x_i, y_i, z_i).

 ISBN: 978-1-61208-398-8

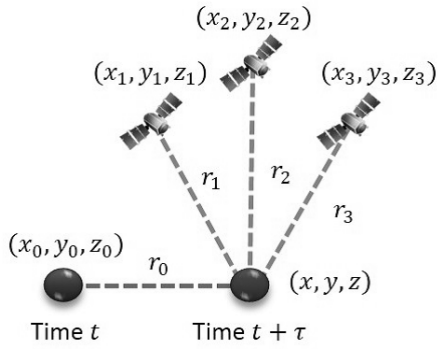


Figure 1. Assumed situation of our research.

By using the orbital information of the satellites which is contained in the satellite signals, the positions of the satellites can be defined. Also, the variables $r_i (i = 1, 2, 3)$ are the distances between the receiver and the satellites. On the other hand, the variables r_0 is the distance between the time t position and the time $t + \tau$ position. Here, we assume that the position was correctly estimated by the adequate satellites at the time t. After traveling, we assume that the observable satellites are decreased at the time $t + \tau$. The number of the observable satellites at the time $t+\tau$ are assumed as three. In this case, the position estimation becomes impossible because of lack of observable satellites.

In this paper, for the purpose of simple explanation, we assume that the number of the observable satellites are four at the time t. The GPS positioning at the time t uses the satellite positions and the distances between the receiver and the satellites. The true distance ρ_i between the i-th satellite and the receiver can be expressed as follows.

$$\rho_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2 + (z_i - z_0)^2} \tag{1}$$

The clock error may be included to the distance between the receiver and the satellite. Therefore, the distance r_i can be expressed by (2).

$$r_i = \rho_i + s \tag{2}$$

Here, the clock error is represented as the parameter s. The unit of clock error s is meter. This equation can be applied to all the observable satellites. As the number of the observable satellites are four at the time t, the following four equations can be obtained.

$$\begin{cases} r_1 = \rho_1 + s \\ r_2 = \rho_2 + s \\ r_3 = \rho_3 + s \\ r_4 = \rho_4 + s \end{cases} \tag{3}$$

By solving (3), it is possible to find out the position of the receiver (x_0, y_0, z_0) and the clock error s [10].

In case of the time $t + \tau$, the above method is not applicable for the position estimation because the number of the observable satellites are three. Therefore, we will estimate the position (x, y, z) by adding the quasi-satellite. That is, we use the previous position (x_0, y_0, z_0) and the distance between the current position and the previous position r_0. The distance r_0 can be represented by (4).

$$r_0 = \sqrt{(x_0 - x)^2 + (y_0 - y)^2 + (z_0 - z)^2} \tag{4}$$

At the time $t + \tau$, we can observe the three satellites. So, we can obtain the three equations (2). By using the three relations based on (2) and (4), we can derive the following equations.

$$\begin{cases} r_1 = \sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2} + s \\ r_2 = \sqrt{(x_2 - x)^2 + (y_2 - y)^2 + (z_2 - z)^2} + s \\ r_3 = \sqrt{(x_3 - x)^2 + (y_3 - y)^2 + (z_3 - z)^2} + s \\ r_0 = \sqrt{(x_0 - x)^2 + (y_0 - y)^2 + (z_0 - z)^2} \end{cases} \tag{5}$$

It is possible to estimate the position (x, y, z) by (5).

The advantage of the proposed method is that we can increase the available satellites by adding (4). In addition, the proposed method only uses the GPS receiver and the distance sensor, no other infrastructure is needed. In our proposal, we need the distance between the current position and the previous position. In recent vehicles, it is easy to measure the traveling distance because most of the vehicles have distance sensors, such as speed pulses.

C. Calculation Process of Position Estimation

Because the simultaneous equation (5) is nonlinear, the solution can be obtained by the sequential approximation that is performed the linearization around the initial value. Here, the procedure of the sequential approximation is shown below. As a notation, subscripts of the right shoulder of the following variables indicate the times of the sequential approximation.

1) we prepare the suitable initial values x^0, y^0, z^0, s^0 about x, y, z, s.
2) By using the receiver position x^0, y^0, z^0 and the clock error s^0, we calculate the distances between the receiver and the satellites.

$$\begin{cases} r_1^0 = \sqrt{(x_1 - x^0)^2 + (y_1 - y^0)^2 + (z_1 - z^0)^2} + s^0 \\ r_2^0 = \sqrt{(x_2 - x^0)^2 + (y_2 - y^0)^2 + (z_2 - z^0)^2} + s^0 \\ r_3^0 = \sqrt{(x_3 - x^0)^2 + (y_3 - y^0)^2 + (z_3 - z^0)^2} + s^0 \\ r_0^0 = \sqrt{(x_0 - x^0)^2 + (y_0 - y^0)^2 + (z_0 - z^0)^2} \end{cases} \tag{6}$$

3) The residual error $\Delta r_i = r_i - r_i^0$ can be determined by using the distance $r_i (i = 0, 1, 2, 3)$ which is actually measured.
4) Since it is possible to approach the correct solution by compensating same amount corresponding to the residual error for x^0, y^0, z^0, s^0, the compensation amount is determined using the partial derivative about x, y, z, s.

$$\begin{array}{rcl} \frac{\partial r_i}{\partial x} & = & -\frac{(x_i - x)}{r_i} \\ \frac{\partial r_i}{\partial y} & = & -\frac{(y_i - y)}{r_i} \\ \frac{\partial r_i}{\partial z} & = & -\frac{(z_i - z)}{r_i} \\ \frac{\partial r_i}{\partial s} & = & \begin{cases} 1 (i = 1, 2, 3) \\ 0 (i = 0) \end{cases} \end{array} \tag{7}$$

From (7), the compensation amount $\Delta x, \Delta y, \Delta z, \Delta s$ to update x^0, y^0, z^0, s^0 can be represented as follows.

$$\begin{cases} \Delta r_1 = \frac{\partial r_1}{\partial x}\Delta x + \frac{\partial r_1}{\partial y}\Delta y + \frac{\partial r_1}{\partial z}\Delta z + \frac{\partial r_1}{\partial s}\Delta s \\ \Delta r_2 = \frac{\partial r_2}{\partial x}\Delta x + \frac{\partial r_2}{\partial y}\Delta y + \frac{\partial r_2}{\partial z}\Delta z + \frac{\partial r_2}{\partial s}\Delta s \\ \Delta r_3 = \frac{\partial r_3}{\partial x}\Delta x + \frac{\partial r_3}{\partial y}\Delta y + \frac{\partial r_3}{\partial z}\Delta z + \frac{\partial r_3}{\partial s}\Delta s \\ \Delta r_0 = \frac{\partial r_0}{\partial x}\Delta x + \frac{\partial r_0}{\partial y}\Delta y + \frac{\partial r_0}{\partial z}\Delta z \end{cases} \tag{8}$$

 ISBN: 978-1-61208-398-8

Here, the simultaneous equation (8) can be represented by the matrix form in order to simplify handling. We define the vectors $\Delta\vec{x} = [\Delta x, \Delta y, \Delta z, \Delta s]^{\mathrm{T}}$ and $\Delta\vec{r} = [\Delta r_1, \Delta r_2, \Delta r_3, \Delta r_0]^{\mathrm{T}}$ (the notation T expresses a transpose) , the equation (8) can be expressed as follow.

$$G\Delta\vec{x} = \Delta\vec{r} \tag{9}$$

Here, the matrix G is usually called as the observation matrix or the design matrix. The matrix G can be expressed as follows.

$$G = \begin{bmatrix} \frac{\partial r_1}{\partial x} & \frac{\partial r_1}{\partial y} & \frac{\partial r_1}{\partial z} & \frac{\partial r_1}{\partial s} \\ \frac{\partial r_2}{\partial x} & \frac{\partial r_2}{\partial y} & \frac{\partial r_2}{\partial z} & \frac{\partial r_2}{\partial s} \\ \frac{\partial r_3}{\partial x} & \frac{\partial r_3}{\partial y} & \frac{\partial r_3}{\partial z} & \frac{\partial r_3}{\partial s} \\ \frac{\partial r_0}{\partial x} & \frac{\partial r_0}{\partial y} & \frac{\partial r_0}{\partial z} & \frac{\partial r_0}{\partial s} \end{bmatrix} = \begin{bmatrix} -\frac{(x_1-x)}{r_1} & -\frac{(y_1-y)}{r_1} & -\frac{(z_1-z)}{r_1} & 1 \\ -\frac{(x_2-x)}{r_2} & -\frac{(y_2-y)}{r_2} & -\frac{(z_2-z)}{r_2} & 1 \\ -\frac{(x_3-x)}{r_3} & -\frac{(y_3-y)}{r_3} & -\frac{(z_3-z)}{r_3} & 1 \\ -\frac{(x_0-x)}{r_0} & -\frac{(y_0-y)}{r_0} & -\frac{(z_0-z)}{r_0} & 0 \end{bmatrix} \tag{10}$$

The compensation amount $\Delta x, \Delta y, \Delta z, \Delta s$ in (8) can be derived by multiplying the inverse matrix of G from the left of (9). Therefore, the compensation amount $\Delta x, \Delta y, \Delta z, \Delta s$ can be determined by solving (11).

$$\Delta\vec{x} = G^{-1}\Delta\vec{r} \tag{11}$$

5) The initial values x^0, y^0, z^0, s^0 are updated by $\Delta x, \Delta y, \Delta z, \Delta s$ as follows.

$$\begin{aligned} x^1 &= x^0 + \Delta x \\ y^1 &= y^0 + \Delta y \\ z^1 &= z^0 + \Delta z \\ s^1 &= s^0 + \Delta s \end{aligned} \tag{12}$$

6) After updating the initial value to x^1, y^1, z^1, s^1, we return to the Procedure 2. These procedures are repeated until $\Delta x, \Delta y, \Delta z, \Delta s$ becomes enough small.

By following the above procedure, our proposed method has the possibility to calculate the receiver's position (x, y, z). In our experience, the solution can converge by repeating several times even if the initial values are $x^0 = y^0 = z^0 = s^0 = 0$.

IV. The Characterization by Field Experiment

A. Setup and Environment

The experiments were conducted in order to evaluate the ability and the effectiveness of the proposed method. In the proposed method, the position estimation is performed while updating the traveling distance and the previous position. The assumed environment is an urban area. There are many buildings in urban areas. The satellites with low elevations tend to be shaded by the buildings. So, we trust only the satellites with high elevations.

In the experiment, we are using a bicycle as the moving vehicle. The bicycle is shown in Figure 2. As shown in Figure 2, a GPS antenna is attached to the loading platform. Also, the cycle computer is attached to the front

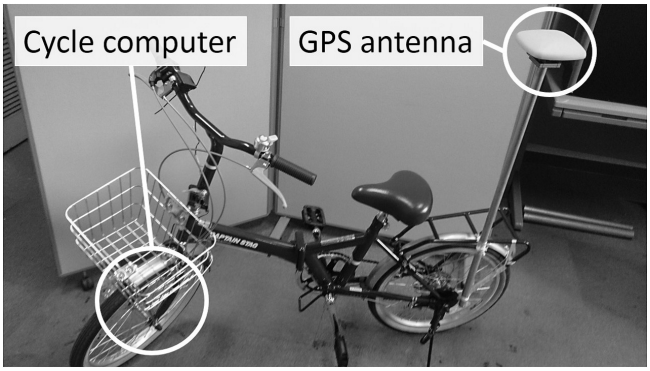


Figure 2. Experimental vehicle.

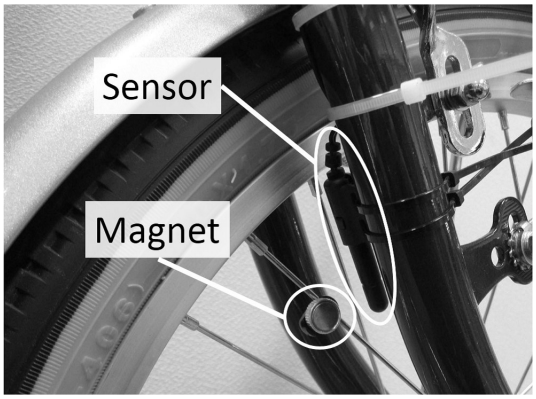


Figure 3. Cycle computer.

wheel. The cycle computer is shown in Figure 3. The cycle computer is a device which senses a magnet mounted on the spokes of the tire to generate a pulse after each rotation. The bicycle also has a data logger system. Each generated pulse is saved in the data logger. The sampling interval of the data logger is 1 ms. The example of the saved pulses is shown in Figure 4. From Figure 4, the cycle computer outputs 0V when the magnet passes the front of the sensor. Except above, the cycle computer usually outputs 2V. We can recognize the rotation of the tire by the pulses. Based on these pulses, the distance r_0 can be measured. The duration between an edge of a pulse and that of the next pulse is equal to the circumference of the tire. The traveling distance r_0 is calculated for each position estimation by using the circumference of the tire.

The experiment has been conducted under an open sky. The distance r_0 had measured while traveling by the bicycle. The position estimation of the proposed method is performed using the three satellites with high elevation and the previous position. For comparison, the positions are also estimated by the conventional method with all the observable satellites.

A total distance of the experimental riding is 100 meters. In the first 20 meters, we rode toward east straightly. Then, we turned right. In the next 80 meters, we rode toward south straightly again. The cycle computer has a function that displays the speed according to the rotational speed of the tire. We kept the speed of the bicycle 10 km/h. From the start to the end, the time is 50 seconds. Table I summarizes the parameters of the experiment environment.

By using the data obtained in the experiment, the position estimation is performed per second. The GPS

 ISBN: 978-1-61208-398-8

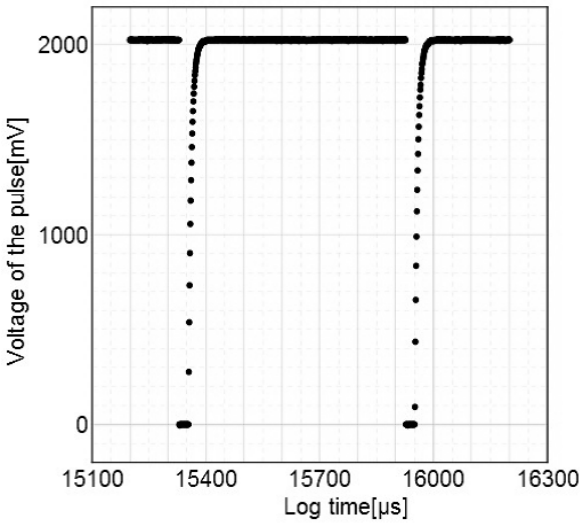


Figure 4. Waveform from the cycle computer.

TABLE I. SPECIFICATIONS OF THE EXPERIMENT ENVIRONMENT

GPS receiver	JAVAD GNSS DELTA-G3T
Number of observable satellites	8 satellites
Total traveling time	50 s
Estimation interval of GPS receiver	1 s
Data logger	EasySync DS1M12
Cycle Computer	CATEYE CC-VL820 VELO 9
Sampling interval of the data logger	1 ms
Circumference of the tire of the bicycle	1.515m

receiver can output the distance between the receiver and the satellites. The output distance includes some errors, such as the ionospheric delay error and the tropospheric delay error (so, the output distance is often called as the pseudo range). The ionospheric delay can be estimated by the transmitted messages from the satellites because the messages have coefficients of equations which are modeled as the ionospheric delay. So, we subtract the modeled ionospheric delay from the output distance. Similarly, we subtract the modeled tropospheric delay from the output distance. We use the remaining distance as the distance r_i For comparison, the position estimation using all satellites was also calculated per second. In the proposed method, the position coordinates of the start is estimated by using the four satellites with high elevation.

B. Position Estimation Results

Figure 5 shows the results of the proposed and conventional method of position estimation. The origin of Figure 5, is the starting point. The positioning results are plotted every second. According to Figure 5, by using the proposed method, the position estimation can be performed even when the number of the observable satellites are three. First, the bicycle is moving towards east, from the start point. After going 20 meters straightly, the direction is changed to the south. Finally, the bicycle stops when the traveling distance becomes 80 meters from the turned point. As we can see, the proposal method can keep estimating the position when the direction of the traveling is changed while traveling.

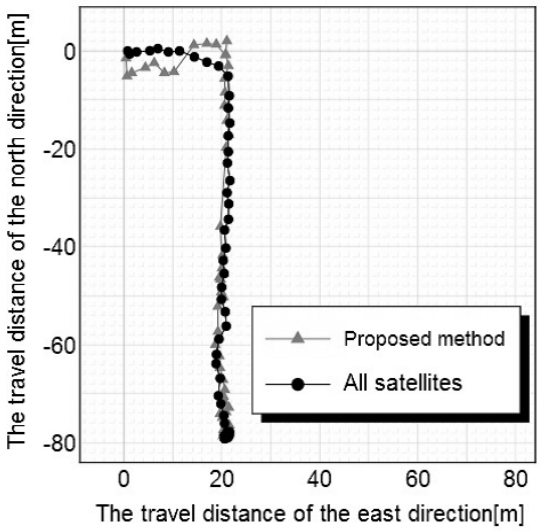


Figure 5. Position estimation results of all satellites and the proposed method.

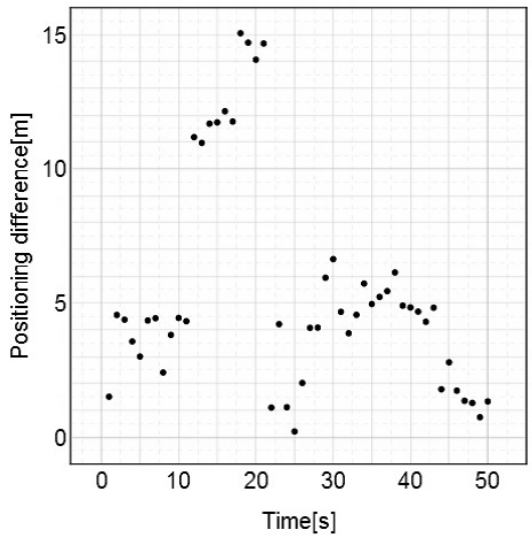


Figure 6. Positioning difference.

In Figure 6, the positioning difference between the estimated position by the proposed method and the position by all the satellites is shown. The positioning difference is defined as Euclidean distance between both the positions. Figure 6 is plotted every second. The total traveling time is 50 seconds. The positioning difference is 5 meters or less until 11 seconds from the start time. In addition, from 22 seconds to 50 seconds, the positioning difference also under 5 meters. Also, the few samples are 7 meters or less. From 12 seconds to 21 seconds, the positioning difference is over 10 meters.

As another viewpoint of discussion, Figure 7 shows the cumulative probability distribution in order to check the distribution of the positioning difference. From Figure 7, 76 percent of the positioning differences are 5 meters or less. The rate of the positioning differences over 10 meters is about 17 percent.

By considering these results, the proposed method is able to estimate the receiver's position by using the previous position and the three satellites with high eleva-

 ISBN: 978-1-61208-398-8

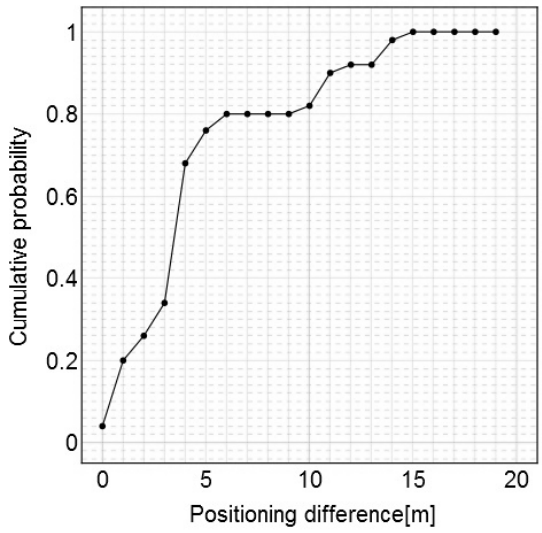


Figure 7. Cumulative probability distribution of the positioning difference.

tion. However, there are some cases when the positioning differences are more than 10 meters. These large differences are a problem which will have to be resolved. One of the above reasons is the satellites constellation. Generally, in case of the four satellites, the good satellites constellation can be presented as follows [11].

- One satellite is near zenith, that is, with high elevation.
- Other three satellites are distributed and surrounded uniformly with low elevations.

In this experiment, three satellites with high elevation have been selected. A better selection has to be consider for the above appropriate satellite constellation. By a better selection, we expect that the proposal can become better.

We now investigate other reasons why the large differences occur. In this paper, the properties were evaluated by comparing the position estimation results of using all satellites. However, the measurement error of the estimation results by all satellites may be included. In oder to investigate in more detail, it is necessary to evaluate the characteristics by comparing the true position with the proposed method.

The proposed process is not much different from the conventional process. In the proposed method, we just use the traveling distance instead of the range from a satellites. So, there is no big difference in calculation time compared to the conventional positioning. We hope that our proposal can be calculated in real-time.

V. Conclusion

In this paper, our proposal was to make position estimation possible in places where the observable satellites are decreased. In order to compensate the decrement, our proposed method assumed the previous own position as the quasi-satellite. For using the previous position as the quasi-satellite, we needed the traveling distance from the previous position. So, we focused on the traveling distance sensor which general vehicles have. By using the distance sensor, we proposed the position estimation method which can keep estimating own position even when the number of the observable satellites becomes low. In order to evaluate the proposed method, the positioning experiment was done using the bicycle. Through the experimental results, possibility of the position estimation and the characteristic of the position estimation were shown when the proposed method is used. Also, there were some cases where the position estimation error was larger than expected. As a future work, we have to consider the reason of the large error and will provide countermeasures.

Acknowledgment

A part of this research is supported by the Cooperative Research Project of Research Institute of Electronics, Shizuoka University and the research promotion of the Murata science foundation.

References

[1] T. H. Dixon, "An introduction to the global positioning system and some geological applications," Reviews of Geophysics, vol. 29, no. 2, 1991, pp. 249-276.

[2] H. Hatano, T. Kitani, M. Fujii, Y. Watanabe, and H. Onishi, "A Helpful Positioning Method with Two Gnss Satellites in Urban Area," IARIA International conference on Mobilie Services, Resources, and Users, 2013, pp. 41-46.

[3] H. Hatano et al., "Positioning Method by Two Gnss Satellites and Distance Sensor in Urban Area," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E98-A, no. 1, 2015, pp. 275-283.

[4] S. Skone, and S. M. Shrestha, "Limitations in DGPS positioning accuracies at low latitudes during solar maximum," Geophysical Research Letters, vol. 29, no. 10, pp. 81-1-81-4, 2002, doi:10.1029/2001GL013854.

[5] Y. Fuke, and E. Krotkov, "Dead Reckoning for a Lunar Rover on Uneven Terrain," IEEE, vol. 1, 1996, pp. 411-416.

[6] H. Kanoh, "Dynamic route planning for car navigation systems using virus genetic algorithms," IOS Press, vol. 11, no. 1, 2007, pp. 65-78.

[7] S. Kim, and J. H. Kim, "Adaptive Fuzzy-Network-Based C-Measure Map-Matching Algorithm for Car Navigation System," IEEE, vol. 48, no. 2, 2002, pp. 432-441.

[8] C. Brenner, and B. Elias, "Extracting landmarks for car navigation systems using existing GIS databases and laser scanning," ISPRS Archivers, vol. XXXIV, part3/W8, 2003, pp. 131-136.

[9] M. S. Braasch, "GPS Receiver Architectures and Measurements," IEEE, vol. 87, no. 1, 1999, pp.48-64.

[10] P. Misra, and P. Enge, "Global Positioning system: signals, measurements, and performance," Ganga-Jamuna Press, 2001.

[11] X. Meng, G. W. Roberts, A. H. Dodson, E. Cosser, J. Barnes, and C. Rizos, "Impact of GPS satellite and pseudolite geometry on structural deformation monitoring: analytical and empirical studies," Journal of Geodesy, vol. 77, no. 12, 2004, pp. 809-822, doi:10.1007/s00190-003-0357-y.

 ISBN: 978-1-61208-398-8

Characterization and Modelling of YouTube Traffic in Mobile Networks

Géza Horváth, Péter Fazekas
Department of Networked Systems and Services
Budapest University of Technology and Economics
Budapest, Hungary
gezah@hit.bme.hu, fazekasp@hit.bme.hu

***Abstract*—Video streaming is one of the most data-intensive applications of today's Mobile Internet and YouTube generates 20% of mobile networks downstream traffic in several regions. YouTube employs the progressive download technique for video playback and therefore its traffic is bursty. We present the characteristics of the most important burst measures of traffic generated by YouTube when accessed via mobile broadband connections. Moreover, we also distinguish the characterization for non-optimized and optimized traffic since mobile operators are using media optimization platforms to effectively deliver video content. In this paper, we present our measurement-based analytical results to derive characterization of the traffic sources. As a result, we propose a generic model and its parameters according to the optimized and non-optimized traffic sources based on the experimental evaluation of the captured YouTube traffic. The proposed traffic models can be used in simulation of future work.**

Keywords-YouTube; video optimization; burstiness; traffic characterization; traffic model.

I. INTRODUCTION

Mobile data traffic is set explode in the upcoming years as consumers add more devices to the mobile networks and operators deploy faster networks. Mobile operators around the world are ramping up deployment of 4G LTE networks while subscribers are consuming more and more data. One of today's most data-intensive applications is video streaming. In this work, we present the characterization and examine the bursty nature of the mobile network traffic generated by one of the most relevant video streaming platform: YouTube.

According to the Sandvine Global Internet Phenomena Report from 2014 in most regions, YouTube is the application responsible for generating the most bandwidth; it accounts for around 20 % of mobile downstream traffic in North America, Europe and Latin America. While Netflix saw growth in the share thanks to the continued rollout of high bitrate Super HD content, in many regions YouTube continues to be the largest single source of Real-Time Entertainment traffic on both fixed and mobile access networks, which makes it the leading source of Internet traffic in the entire world. In North America 17.61% of the total downstream traffic is generated via YouTube, while in Europe it is 19.27% [1].

YouTube employs the progressive download technique; its video content is delivered by a regular HTTP web server rather than a streaming server. Video delivered using this technique is stored on the viewer's hard drive as it's received, and then it's played from the hard drive; it enables video playback before the content download is completely finished [6]. It also uses the HTTP/TCP (Hypertext Transfer Protocol/Transmission Control Protocol) platform to deliver data, which further distinguishes it from traditional media streaming [4].

The present paper aims at four main objectives: (1) to highlight the YouTube service traffic characterization when accessed via mobile broadband connection; (2) to present proper measures for catching the bursty nature of YouTube traffic; (3) to distinguish the effect of media optimization platform used in mobile networks; and (4) to propose traffic models for non-optimized and optimized YouTube traffic accessed via mobile broadband connection.

The rest of the article is organized as follows: Section 2 provides an overview of the traffic sources and the experimental framework used during the analysis and the capture method itself; Section 3 provides a summary of the most important burst and correlation measures and their numerical evaluation on the captured traces; Section 4 presents the analysis of the experimental results of the main characteristics of YouTube traffic accessed via 3G mobile network; Section 5 provides the traffic model state machine, its parameters and its mode of operation via algorithm; finally, Section 6 presents the main conclusions.

II. TRAFFIC SOURCES

A. Experimental framework

In this section, we describe the experimental framework used to collect traces of data traffic generated by YouTube accessed via 3G mobile network. The framework is composed of a notebook connected to a mobile service provider's network via USB stick on the move. The modem is able to handle downstream traffic up to 42 Mbps via DC-HSPA+ (Dual Carrier-High Speed Packet Access) and 21 Mbps via HSPA+ access. Its upstream capability is 5.76 Mbps. The used mobile technology was at least HSPA+ in 98.4% of the samples, the remaining part was EDGE (Enhanced Data Rates for GSM Evolution). During pilot tests it was verified that neither the CPU (Central Processing Unit) nor the memory of the notebook impeded the normal playback of the video clips.

A playback monitor tool has been built with the main objective of analyzing YouTube traffic and collect available information about the videos. The tool includes a web application using the YouTube player API (Application

 ISBN: 978-1-61208-398-8

Programming Interface) via JavaScript [11]. It is able to play videos sequentially based on *video_id* list and collect all the available information like duration, total bytes into a log file. The web application was stored on a public web server and was accessed from the notebook via Chrome browser 39.0. The other part of the framework is Microsoft Network Monitor 3.4 software installed on the notebook. However, Wireshark is more popular for trace collection, we had to swap it because it was not able to capture on mobile broadband interfaces. On the other hand the collected traces were compatible with Wireshark as well, so we could use its advantages in traffic analysis. It was enough to capture only the first 68 bytes of each package, with the help of the headers we were able to reproduce the traffic itself [9] [10] [12].

B. Media Optimization in Mobile Networks

Video optimization refers to a set of technologies used by mobile service providers to improve the consumers viewing experience by reducing video start times or re-buffering events. The process also aims to reduce the amount of network bandwidth consumed by video sessions. While optimization technology can be applied to videos played on a variety of media-consuming devices, the costliness of mobile streaming and increase in mobile video viewers has created a very high demand for optimization solutions among mobile service providers [8].

Video optimization techniques used to improve network performance and subscriber experience include:

- Caching — local copies of video are stored for successive access
- Pacing — data is delivered "just in time" rather than all at once (and frequently abandoned)
- Transrating — changing the frame rate
- Transcoding — recoding the video codec for lower bitrate
- Enhancing TCP — for mobile handling to minimize back and forth signaling
- Compressing — data reduction of the content

Lossless Media Optimization offers a low-cost media optimization entry-point for operators. It adjusts the video transmission rate to match the actual video play rate, without changing the video content. Instead of clogging the network at the start of the video, the lossless Just-In-Time (JIT) technique "spreads" the video delivery over the entire video play time. With JIT, videos download as fast as needed, rather than as fast as possible, taking additional advantage over the fact that most videos are not watched to the end [8].

Lossy Media Optimization reduces the amount of data transmitted over the wireless network. Operators are using static data reduction or dynamic bandwidth shaping, which are performed in combination with Just-In-Time (JIT) lossless optimization. To achieve the best response time, lossy optimization is performed "on-the-fly". The data reduction of the media is a function of the media quality. The higher the data reduction of the content, the lower is the quality of the media provided [8]. Within the measured mobile network Lossy Media Optimization is in use. Figure 1 depicts the high-level architecture of the platform.

We have two types of traffic in the mobile network. In the first scenario video clips were delivered through the aforementioned media optimization platform, in the second scenario we collected traces without that. The mobile network was configured with two different Access Point Names to differentiate between the two scenarios. The only thing we had to do is to set the APN in use on the notebook before connecting to service provider.

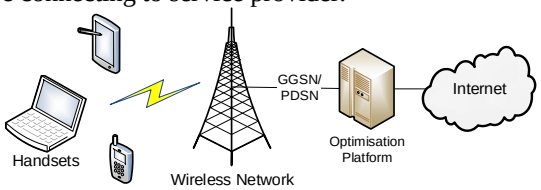


Figure 1. Architecture of Opimisation Platform

C. Collected traces

A trace set of 95 video clip downloads was collected with the recently described experimental framework. This set has been collected to understand the main traffic characteristics of YouTube traffic accessed via mobile network. All the video clips has been downloaded in their default format [4][5]. YouTube uses the MP4 container for high-definition (HD) clips and uses the Flash Video (FLV) as the default format for the majority of non-HD clips. YouTube adapts all the uploaded clips to the aforementioned formats before posting [4].

The trace set was captured two times as follows. We can assume that there was no bottleneck in the mobile networks backhaul or core network.

- SHAPED: This set has been collected using the recently mentioned mobile networks Lossy Media Optimization function; APN is set accordingly. Traffic was affected by the YouTube server, the Media Optimization platform and the access technology, radio conditions of the used mobile network.
- UNSHAPED: This set has been collected without using the recently mentioned mobile networks Lossy Media Optimization function; APN is set accordingly. Traffic was affected by the YouTube server and the access technology, radio conditions of the used mobile network.

III. BURST AND CORRELATION MEASURES

A simple class of burstiness measures takes only the first-order properties into account; they are each a function of the marginal distribution only of interarrival times (with resolution of 0.001 s). These measures can be taken into consideration as various characteristics of the marginal distribution of the inter-arrival time. The possible set of properties are the moments of that distribution [2].

More complex measures are utilizing the second-order properties of the traffic; they do take account of temporal dependence in traffic. From this class indices of dispersion is one of the most well-known methods. It includes the correlation properties of the traffic and can be very informative [2].

 ISBN: 978-1-61208-398-8

A. Measures based on the first order properties

One of the mostly used measures is the peak to mean ratio (PMR). Peak is defined in this paper as inter-arrival time between the two closest arrivals. In case of UNSHAPED traffic peak may be very high and likely to correspond to two arrivals in consecutive slots in practice. In the case of SHAPED traffic, the peak tells more about the shaper parameters than the traffic itself [3].

TABLE I. FIRST ORDER PROPERTIES

	PMR	***SCV***	m_3
UNSHAPED	0.00004349	281.555	25.265
SHAPED	0.00008560	254.040	24.411

Another widely used measure is the squared coefficient of variation (SCV) of the inter-arrival times. It includes information from the first two moments and is defined $C^2(X) = Var(X)/E^2(X)$ where X is the inter-arrival time [3]. As shown in Table I, when comparing the values of the SHAPED and UNSHAPED traffic, it is higher for the UNSHAPED indicating it is burstier because it includes sustained higher intensities. Also, the lower value of the SCV indicates a smoother process.

The SCV takes into account only the set of the inter-arrival times, the order of that is passed by. However, burstiness is mainly caused by two factors [3]: the distribution and especially the tail of the inter-arrival times and the correlation between them. The squared coefficient of variation from the first-order measures class takes into account only the inter-arrival distribution.

Suppose that X is a real-valued random variable. The variance of X is the second moment of X, and measures the spread of the distribution of X. The third and fourth moments of X also measure interesting features of the distribution. The third moment measures skewness, the lack of symmetry, while the fourth moment measures kurtosis, the degree to which the distribution is peaked.

Higher moments can also tell useful information about the traffic characteristics. For example, two traffic with same first two moments but different third moment can produce very different queueing behavior. The third moment tells about the long inter-arrivals. Although inter-arrival times are bounded from above for both SHAPED and UNSHAPED traffic, higher m_3 for UNSHAPED traffic means it has higher inter-arrival times.

B. Measures based on the second order properties

Indices of dispersion measures are useful because they show the traffic variability over different scales and they can capture the correlation structure. Two indices of dispersion measures are widely used: the index of dispersion for intervals (IDI) is related to the sequence of inter-arrivals; the index of dispersion for counts (IDC) is related to the sequence of counts of arrivals in consecutive time units [2].

The IDI is defined for a stationary inter-arrival sequence X as follows:

$$IDI = \frac{var(X_{i+1}+\cdots+X_{i+n})}{nE^2(X)} = C_j^2\left(1 + 2\sum_{j=1}^{n-1}\left(1-\frac{j}{n}\right)\rho_j\right) \quad (1)$$

$$n = 1, 2, \ldots$$

$$C_j^2 = \frac{var(X)}{E^2(X)}$$

$$\rho_n = cov(X_i, X_{i+n})/var(X)$$

In the definition, the sum of k consecutive inter-arrivals is taken. In the case of bursty process, the sort and long inter-arrivals are grouped together, and it causes and it causes the IDI to increase with increasing k. In fact, the increase or decrease in the IDI graph is directly related to the correlation of the inter-arrival sequence.

The IDC for a stationary process is defined as

$$IDC = \frac{V(t)}{E(t)} = \frac{V(t)}{tm} \quad (2)$$

where V(t) and E(t) are the variance and the expected number of the arrivals in an interval of length t, and E(t)=tm, where m is the mean intensity of arrivals. The IDC shows the variability of a process over different time-scales.

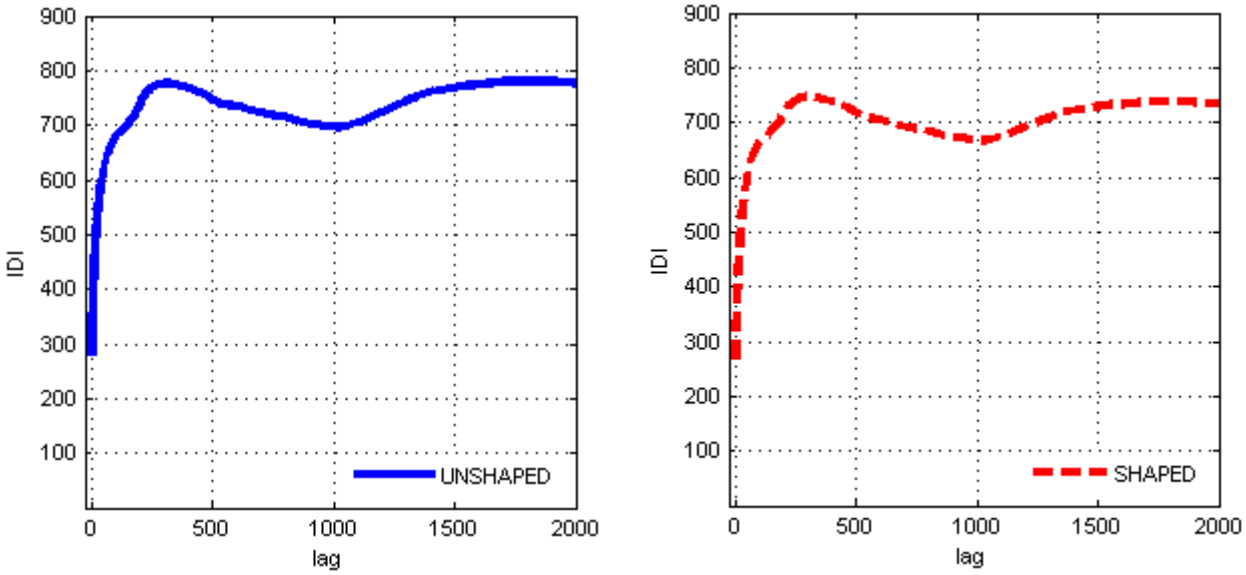


Figure 2. IDI Graph of traffic sources: (a) UNSHAPED, (b) SHAPED

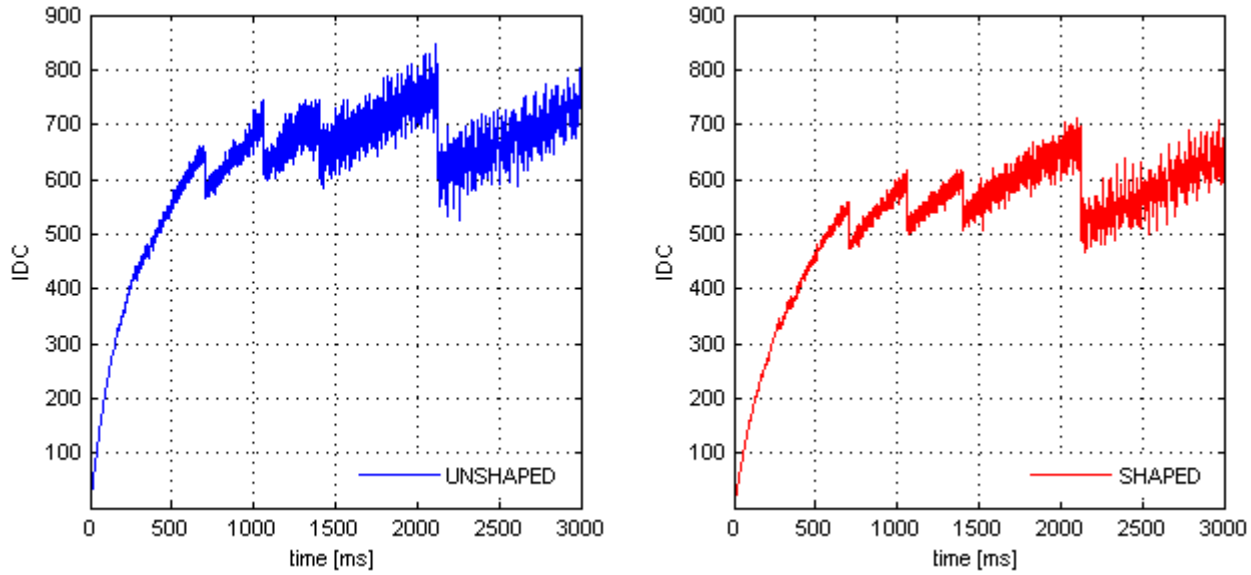


Figure 3. IDC Graph of traffic sources: (a) UNSHAPED, (b) SHAPED

As shown in Figure 2 and Figure 3, in case of UNSHAPED and SHAPED traffic the IDI and IDC curves both increase. Together they imply that the low burstiness in short scales increase over higher scales due to positive correlation. The quickly increasing curves and the high value at infinity imply that these are very bursty sources. From Figure 3, it can be seen, the SHAPED traffic has lower squared coefficient variation (start of IDI curve) and a bit lower IDI and IDC curves than UNSHAPED meaning lower burstiness of the traffic source.

To have accurate IDC curves, the maximal block size (t) should not exceed 10% of the sample size. Using non-overlapping blocks with size *t* we need at least 10 values in a block to calculate accurate variance. From Figure 3 it can be seen that increasing block size *t* implies more inaccurate IDC curves.

IV. ANALYTICAL RESULTS

This section introduces the experimental results obtained to evaluate the traffic generated by YouTube captured via mobile broadband access. On the basis of the information provided by our experimental framework we depict the progressive download of a video clip, which belong to trace set UNSHAPED, as an example. Figure 4 plots the time evaluation of the instantaneous amount of data received by the player at the beginning of the download.

A. *Initial burst*

A video clip download commences with a significant burst of data, later the receiving data rate of the client's player is considerably reduced (see Figure 4 (a)). Initial burst is identified in each trace by determining the slope change in the accumulated data received by the player between the initial burst and the throttling phase. To eliminate the effect

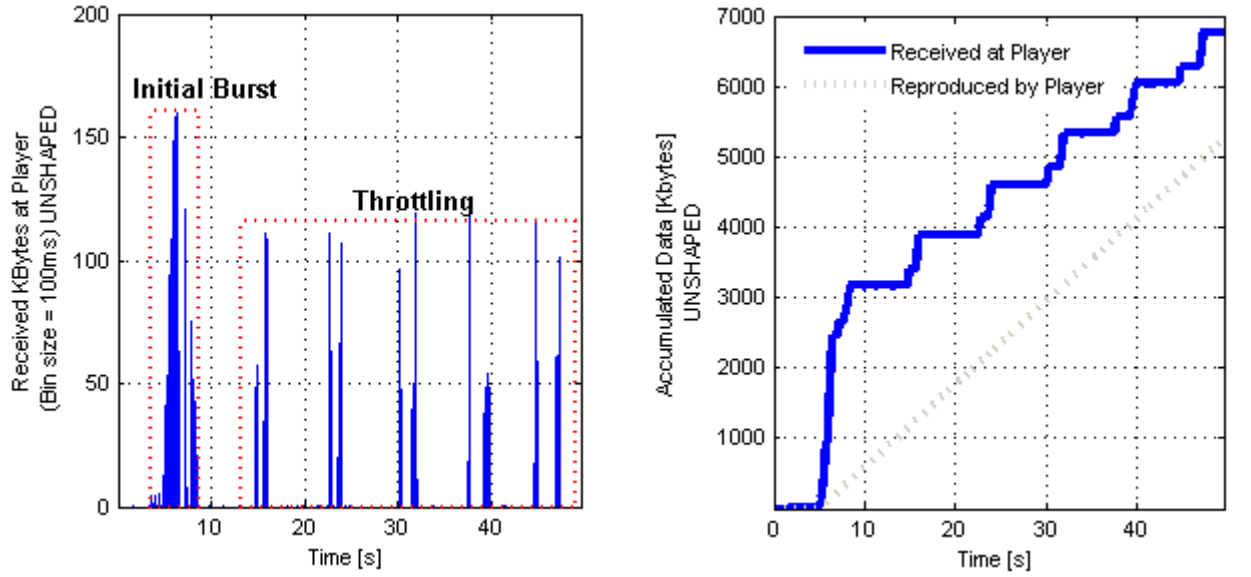


Figure 4. Examples received data at player: (a) instananeous, (b) accumulated

 ISBN: 978-1-61208-398-8

of temporary slop changes caused by network bandwidth fluctuations, the accumulated data have been filtered with a 400 ms simple moving average. A slope approximation sequence computed as the difference between consecutive samples of the filtered series. Then, the maximum slop after the initial burst, during the throttling phase is computed by considering only the last 20 s of the trace [4]. The observed end of the initial burst is measured as the last instant of the trace when two consecutive samples of the slope approximation sequence surpass the maximum slope of the throttling phase.

Figure 5 (a) depicts the CDF of the amount of data (measured in seconds of video data) downloaded until the observed end of the initial burst for all downloads of the traces. The results show that the majority of the measured sizes amount to approximately 52 s and 65 s of the video data. For the remaining downloads, the empirical measurements of their initial burst slightly differ from the above mentioned two values, which is caused by short fluctuations in the mobile network's available bandwidth that affect the empirical estimation.

Our examination highlighted that majority of the collected traces from UNSHAPED and SHAPED traffic sources has the same initial burst size measures in video seconds. However, in case of SHAPED traffic the deviation is a bit higher. Only the download duration of the initial burst is different: it is significantly higher in case of UNSHAPED traffic than SHAPED. It is caused by the traffic shaping function of the aforementioned media optimization platform, limiting the available bandwidth for a given video clip, therefore helping the network not be overloaded.

It should be noted that in opposite with the results of [4] setup parameter burst is not sent via the HTTP request by the YouTube client anymore. Our experimental result show that the initial burst size is not limited to 40 s but can take on different values, higher than earlier.

B. *Throttling algorithm*

We continue our discussion of the experimental analysis by focusing on the traffic received by the player after the initial seconds of a progressive download. As shown in Figure 4 (b), after the initial burst, the slope of the accumulated received data at the player was reduced because of a decrease in the receiving data rate. Figure 4 (b) shows that after the initial burst the steepness of the slope remains approximately constant until the download is completed. It is caused by the server, which throttles down the traffic generation rate increasing the total time required to complete file download. From the collected traces throttling factor can be calculated easily:

$$Throttling\ factor = \frac{Total\ data\ rate}{Throttling\ data\ rate} \quad (3)$$

From the results of Figure 4 it can be concluded that after the initial burst, the media server throttles down the traffic generation rate, thereby avoiding transferring the data at the maximum available bandwidth. A throttling algorithm is applied with a throttling factor of 0.92 of the video total data rate. Figure 6 also depicts that there is no difference in throttling factor of UNSHAPED and SHAPED traffic sources.

This throttling procedure is also used in other platforms. It saves bandwidth of media files that might not be played to the end [7]. Additionally, it prevents congestion both at the server and the network because the data transfer is not performed at the Internet's maximum available bandwidth.

C. *Chunk size*

In the previous section we highlighted that traffic generation rate is constant during the throttling period. If we magnify Figure 4 (a), it is clearly visible that traffic consists of small chunks. Figure 4 shows that during the throttling phase, the pattern of reception of data alternates between the reception of data chunks and short periods without packets.

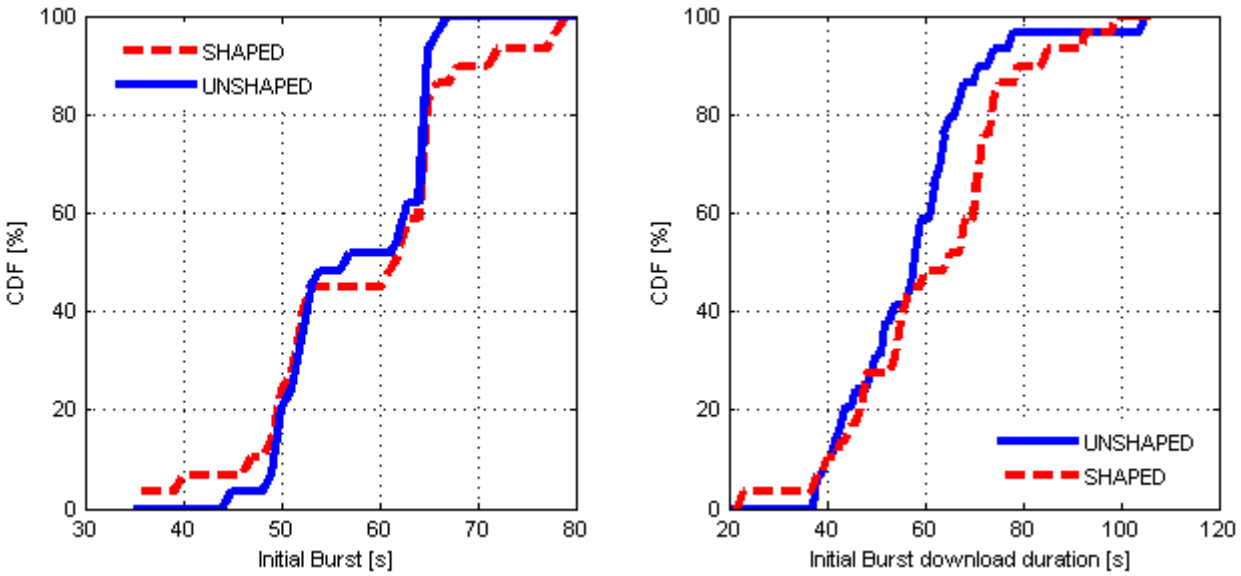


Figure 5. Initial burst measures: (a) Size in video seconds, (b) Download duration

 ISBN: 978-1-61208-398-8

To further analyze this characteristic, all video clips of trace set UNSHAPED and SHAPED were postprocessed. It eliminates the initial burst of each download and additionally, groups packets into chunks so that two consecutive packets belong to the same chunk if the difference between their arrival times does not exceed a given time threshold. If the difference is longer than the time threshold, the two consecutive packets are assumed to belong to different chunks. Thus, the size of a chunk can be calculated simply by aggregating the size of the payloads of all of its TCP packets. The time threshold used to decide if two consecutive packets belong to the same chunk is selected to be 200 ms.

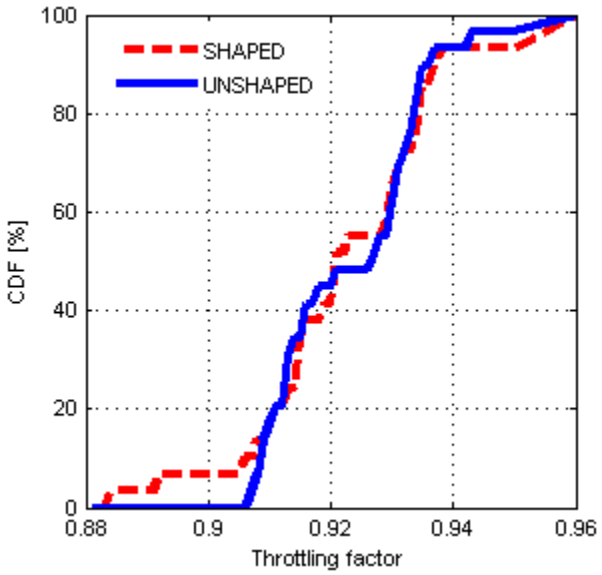


Figure 6. Throttling factor

From the empirically measured chunk sizes, we observe that in case of audio majority of the measured chunk sizes are equal 240 KB. Marginally, chunks with size of 182 KB were also found. We cannot observe significant difference between UNSHAPED and SHAPED traffic sources. In case of video we observe chunks with sizes between 400 and 1500 KB. Very small chunks (40-52 bytes) are ignored since these are identified as ACK messages in the TCP sessions.

V. TRAFFIC MODEL

On the basis of the experiments presented in previous sections, a common synthetic model of the UNSHAPED and SHAPED traffic sources are proposed by means of a basic state machine (see Figure 7.) with pseudo-code describing its mode of operation (see Figure 8.). The algorithm provides the time instants and burst sizes in bytes to send on the TCP layer. The two examined traffic sources are distinguished with the parameters of the state machine as follows.

TABLE II. . PARAMETERS OF INITIAL BURST LENGTH DISTRIBUTION

	a	*b*	*μ1*	*σ1*	*μ2*	*σ2*
UNSHAPED	51.72	48.28	50.54	3.77	64.29	1.32
SHAPED	44.83	55.17	48.24	5.62	66.11	5.17

The generic traffic model consists of two states: an initial burst and throttling state. At the start of a video playback the algorithm needs d (video total duration in seconds) and s (video total size in bytes) as input parameters.

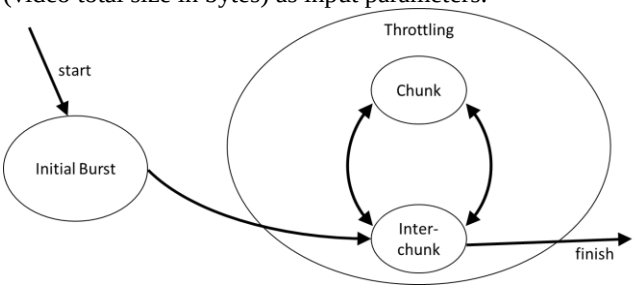


Figure 7. Generic traffic model

To set up the initial burst state first we need to calculate the size of the burst to send in bytes. To achieve this, we recall that distribution of d_{ib} (initial burst size in video seconds) shows two distinct values with high probability. According to Equation (4) we use the weighted sum of two distinct normal distributions as approximation. Parameters of the distribution formula of d_{ib} can be obtained from Table II. The s_{ib} parameter (size of the initial burst in bytes) can be calculated based on d_{ib} easily with Equation (5).

$$d_{ib} = a * \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + b * \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \quad (4)$$

$$s_{ib} = \frac{d_{ib}}{d} * s\ [bytes] \quad (5)$$

While having b available bandwidth, in the initial burst state the algorithm has to send the first s_{ib} bytes of the video to the player with b available bandwidth. Algorithm maintains a variable $s_{remaining}$, which contains the bytes still has to send to the player out of s. Figure 8. presents to pseudo code of each state of the state machine.

After sending an initial burst with s_{ib} bytes, in the throttling state the procedure write blocks of *cs* (chunk size in bytes) of data into the TCP socket with a period controlled by the *tf* (throttling factor). Chunks are generated based on identified chunk sizes.

```
//initial burst
While s_remaining > s - s_ib
    Send initial burst with b bandwidth;
Endwhile;

//throttling state
While s_remaining > 0
    Send chunk with size cs
    Sleep for cs / [(s/d) * tf] seconds
Endwhile;
```

Figure 8. Pseudo-code of the states

We have compared the download rates from the original YouTube and synthetic model traces for every video clip. For the comparison, the instantaneous relative error of the accumulated amount of data has been computed at every sampling instant *n* as:

$$\varepsilon[n] = \frac{\left|\hat{A}[n] - A[n]\right|}{A[n]}$$

where $\varepsilon[n]$ denotes the instantaneous relative error, $A[n]$ represents the amount of the accumulated data received by the player's buffer in the case of the download from the original YouTube server and $\hat{A}[n]$ represents the amount of the accumulated data from synthetic model. It has to be noted that period between consecutive samples of the discrete-time sequence $A[n]$ and $\hat{A}[n]$ was set to 100 ms. Finally the 90th percentile of the discrete-time sequence $\varepsilon[n]$ has been computed and denoted as $\hat{\varepsilon}$. The results show that the relative error $\hat{\varepsilon}$ does not exceed 8%.

VI. CONCLUSION

In this paper we described the characteristics of YouTube traffic from the viewpoint of mobile broadband access. It is very valuable for predicting the video quality perceived by end-users and enhancing network design. The characterization is based on our executed experiments.

The present results have shown, that YouTube traffic has a bursty nature when accessing via mobile broadband connection. It has been also identified, that media optimization has its effect on bursty YouTube traffic: using lossy media optimization with just-in-time delivery function can visibly decrease the burstiness of the traffic. We have verified this via first order properties like SCV, PMR and m_3, and second order properties like IDI and IDC.

We have depicted the differences between UNSHAPED and SHAPED traffic sources: we have presented parameters of initial burst, throttling factor and chunk size for both traffic sources based on our experiments. It was also justified that SHAPED traffic source has the same amount of bytes in the initial burst as UNSHAPED, but is consumes less bandwidth, it takes longer for SHAPED traffic to download the initial burst. It was highlighted that initial burst size parameter is not sent via the HTTP request by the YouTube client anymore; based on our experimental result it is not limited to 40 s.

We proposed a generic traffic model for the examined traffic sources and we also presented its parameters for UNSHAPED and SHAPED YouTube videos. The model is given with its formulas and can be easily implemented in network simulation tools to evaluate service performance and end-user quality. In future work we plan to extend our model with the differentiation between audio and video streams.

REFERENCES

[1] Sandvine Global Internet Phenomena Report 1H 2014 [Cited 2014 Oct 22]. Available from: https://www.sandvine.com/downloads/general/global-internet-phenomena/2014/1h-2014-global-internet-phenomena-report.pdf. May 15, 2014.

[2] S. Molnár, Gy. Miklós, "On Burst And Correlation Structure of Teletraffic Models (extended version)" 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, 21-23 July 1997, West Yorkshire, U.K.

[3] V. S. Frost, B. Melamed, "Traffic Modelling For Telecommunications Networks", IEEE Communications Magazine, March, 1994

[4] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Analysis and modelling of YouTube traffic", Transactions on Emerging Telecommunications Technologies, 2012

[5] Characterization of trace sets T1 and T2. [Cited 2014 Sept 1]. Available from: http://dtstc.ugr.es/tl/downloads/set_t1_t2.csv

[6] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: a view from the edge", In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, 2007, DOI: 10.1145/1298306.1298310.

[7] Microsoft Corporation. IIS Media Services. [Cited 2014 February 27]. Available from: http://technet.microsoft.com/en-us/library/ee729229(WS.10).aspx. June 10, 2010.

[8] Citrix ByteMobile. Applying Adaptive Traffic Management: Improving Network Capacity and the Subscriber Experience. [Cited 2014 October 10.]. Available from: https://www.citrix.com/content/dam/citrix/en_us/documents/products-solutions/applying-adaptive-traffic-management-improving-network-capacity-and-the-subscriber-experience.pdf. 2013.

[9] Wireshark Corporation. Wireshark network protocol analyizer. [Cited 2014. October 10.]. Available from: https://www.wireshark.org/

[10] Microsoft Corporation. Microsoft Network Monitor 3.4 [Cited 2014. October 10.]. Available from: http://www.microsoft.com/en-us/download/details.aspx?id=4865

[11] Youtube Corporation. YouTube APIs and tools.[cited 2014 April 25]. Available from: http://code.google.com/intl/en-US/apis/youtube/overview.html.

[12] G. Horváth, "End-to-end QoS Management Across LTE Networks", In the proceedings of SoftCOM Conference, 2013 DOI: 10.1109/SoftCOM.2013.6671871

 ISBN: 978-1-61208-398-8

Implementation Design of UPCON-based Traffic Control Functions working with vEPC

Megumi Shibuya, Atsuo Tachibana and Teruyuki Hasegawa
KDDI R&D Laboratories, Inc.
Saitama, JAPAN
e-mail: {shibuya, tachi, teru}@kddilabs.jp

Abstract—**To resolve the Radio Access Network (RAN) congestion issue, 3GPP is standardizing a mechanism named User Plane Congestion Management (UPCON) which notifies congestion status information on RANs to Evolved Packet Core (EPC) or behind EPC for efficient traffic control of Long Term Evolution (LTE). This paper presents the implementation design of the traffic control functions (we call them "TCFs") working with virtualized EPC (vEPC). TCF can control traffic in accordance with congestion status information on RANs, which is supposed to be notified by Evolved Node B (eNB) to EPC and its behind systems based on UPCON. We implemented the proposed system as virtual machines working with commercial vEPC software, which includes EPC functions, such as Serving/Packet data network GateWay (S/P-GW) and Policy and Charging Rules Function (PCRF). Through the experimental system, we evaluated the feasibility of TCFs. In addition, we discuss the applicability of TCF to Mobile Network Operators (MNOs).**

Keywords - UPCON; Congestion Control; vEPC; Traffic Control; DIAMETER

I. INTRODUCTION

Since introducing the Long Term Evolution (LTE), the Radio Access Network (RAN) capacity of cellular networks has been expanding dramatically. On the other hand, rich applications, such as high-definition audio/video streaming, image sharing, and online-games have been wide-spreading by which mobile users download rich content via cellular networks. So the traffic volume per user has been increasing dramatically. According to a forecast in [1], from 2013 to 2018, the number of smartphone devices is expected to grow at 18% compound annual growth rate (CAGR) while the mobile data traffic volume is expected to grow at 63% CAGR, i.e., 3.5 times larger. In other words, the growth in traffic volume per user outpaces the growth in the number of devices. Therefore, RAN congestion still remains as a critical issue even if LTE migration is going well [2].

To resolve this issue, the following two solutions easily come up with. First, expanding RAN bandwidth capacity, e.g., by small sizing cells and/or using higher radio frequency is an intuitive and essential solution. However, it requires some processes consuming much time and cost, such as redesign of cells, which may be required again and again corresponding to traffic changes. Second, applying traffic control to congested cells is another solution that just shares the limited bandwidth capacity more reasonably. Although the total capacity does not change, it can improve each user's Quality of Experience (QoE) by preferentially handling critical communications from the other ones and it can more flexibly cope with congestion corresponding to traffic changes. Therefore we focus on the latter solution in this paper.

As for the concrete solutions of traffic control in response to congestion status information on RANs, we roughly categorize them into the following three types. First is that in transport level, e.g., TCP controls the traffic for avoiding congestion in wireless network with lower layer information, where TCP adjusts to the most suitable transmission rate dynamically in accordance with channel status [3]. Second is that based on packet-level QoS scheduling, such as priority queuing. It can provide higher priority communications to users by prioritizing packets generated from specific applications related to the communications [4]. Third is that so-called traffic offload, where (all or a part of) the traffic is redirected to surrounding cells or other types of networks to reduce the traffic in the target cell. For example, Coordinated Multi-Point (CoMP) is standardized by 3GPP release 11 [5], where the congestion level on each Evolved Node B (eNB) is monitored in real time to avoid congestion by load balancing the traffic with surrounding eNBs. In regard to the other types of networks, Wi-Fi or wired network (e.g., FTTH) is used in general.

From the fairness viewpoint between users in a congested cell, traffic control mitigating congestion should not be applied on a *per-flow* (e.g., TCP connection) basis but per *user* (e.g., UE: User Equipment) basis. Moreover, when a UE attaches Wi-Fi, all the traffic from/to the UE had better be offloaded to Wi-Fi, which is another reason for applying per user basis traffic control. In addition, since such other types of networks are being accommodated in Evolved Packet Core (EPC) or behind EPC, traffic control should work in or behind EPC. Hence, a certain mechanism is required by which EPC (or the systems behind it) knows the congestion status information on RANs. 3GPP is now standardizing such a mechanism named "User Plane Congestion Management (UPCON)" [6]. Some use cases of UPCON are proposed in [7][8], such as controlling the traffic at peak time during commuting time, and giving priority accesses for premium users who pay a premium (and expensive) fee. However, feasibility studies on their designs and implementations have not yet been carried out enough.

In this paper, we propose an implementation design of the traffic control functions (hereinafter called "TCFs")

 ISBN: 978-1-61208-398-8

working with virtualized EPC (vEPC). TCF can control the traffic in accordance with congestion status information on RANs (hereinafter called "RAN congestion status"), which is supposed to be notified by eNB to EPC and its behind systems based on UPCON. Specifically, our implementation design covers the executions of multiple (and different types of) TCFs independently based on RAN congestion status, which can be retrieved from each TCF with a reasonable overhead. Furthermore, we implemented the experimental system as virtual machines working with a commercial vEPC software [9], which includes EPC functions, such as Serving/Packet data network GateWay (S/P-GW) and Policy and Charging Rules Function (PCRF) [10]. We also evaluated the feasibility of TCFs through the system.

This paper is organized as follows. Section II summarizes the framework of LTE system and UPCON. Section III explains the proposed implementation design of the traffic control functions (TCFs). In Section IV, we evaluate the feasibility of the proposed system, then discuss the applicability of TCF to Mobile Network Operators (MNOs) in Section V. Finally, we conclude the work in Section VI.

II. FRAMEWORK OF LTE SYSTEM AND UPCON

Figure 1 shows an example of LTE system configuration. LTE system consists of UE, eNB, and EPC. EPC includes Mobility Management Entity (MME) and PCRF at Control Plane (C-Plane), and includes Serving data network GateWay (S-GW) and Packet data network GateWay (P-GW) at User Plane (U-Plane). PCRF sets policy rules related to QoS and charging as QoS Class Identifiers (QCIs) and sends them to S/P-GW. Traffic control is conducted based on QCIs. Behind EPC, Application Function (AF) exists in the Packet Data Network (PDN) (e.g., Internet, IMS). Rx interface [11] based on DIAMETER protocol [12] is prescribed between PCRF and AF.

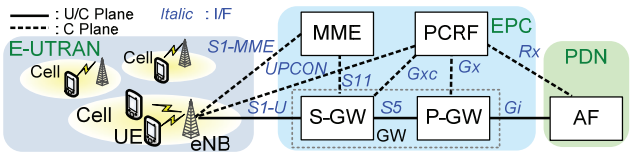


Figure 1. Example of LTE system configuration.

Recently, although the traffic can be controlled in accordance with QCIs set at PCRF in EPC, there has been no standardized mechanism for EPC to grasp the RAN congestion status. In addition, since other types of networks for traffic offload are being accommodated in or behind EPC, traffic control should work behind the accommodation point. Hence, the mechanism called "UPCON" by which the RAN congestion status is notified to EPC or its behind is now being standardized in 3GPP.

In order to further discuss the concrete implementation issues, we suppose that RAN congestion statuses are notified by each eNB, which forms a cell as shown in Fig. 1, and we set the TCFs at AF located behind EPC. In addition, we assume that RAN congestion levels are defined based on each MNO's service operation policy. Since the RAN congestion will occur at the network between an eNB and UEs, i.e., inside a cell, RAN congestion status will vary on a per-cell basis (hereinafter called "*per-cell basis*").

III. TRAFFIC CONTROL FUNCTIONS

As discussed in Section I, TCF should be applied on a per-user basis, while RAN congestion status will vary on a per-cell basis as described in Section II. Thus, we carefully established our design principals of TCF implementation considering how TCF handles relationships between UE and eNB with a reasonable overhead in terms of the data volume and processing time. It should be noted as beyond the scope of this paper that how PCRF obtains RAN congestion statuses from eNBs.

A. Design Principles

First, we established the following design principles.

1. RAN congestion occurs on a per-cell basis. In contrast, TCF assumes that the traffic is controlled on a per-user basis (hereinafter called "*per-UE basis*"), which is practically identified by UE IP addresses. Namely, TCF needs to grasp the RAN congestion level on a per-UE basis. We compared the following notification methods of RAN congestion statuses from PCRF to TCF;
   - Method-1: notify the RAN congestion statuses and information of all the visited UEs in an eNB on a per-cell basis.
   - Method-2: notify the RAN congestion status of each UE accommodated in the TCF, i.e., on a per-UE basis.

   We selected Method-2 because, 1) all the visited UEs in the eNB do not always use the TCF, and 2) when multiple TCFs exist and different UEs are accommodated independently, some overheads on redundant extractions of UEs in the same eNB will occur in all the TCFs if per-cell based notification is applied.
2. RAN congestion statuses are exchanged between PCRF and TCF in accordance with Rx interface. So as to be available for multiple TCFs, a unique Rx session ID, e.g., which is created from the UE's IP address and TCF identifier, is assigned to each UE and TCF pair as a unit.
3. Since each TCF will accommodate different UEs when multiple TCFs exist, a congestion status database (hereinafter called "CDB") for maintaining the congestion statuses of UEs is created in each TCF to keep the size of each CDB as small as possible. The details of CDB are explained in Section III *B*.
4. We consider two types of the notification trigger about RAN congestion status changes in eNBs;
   - Trigger-A: TCF requests RAN congestion status to PCRF periodically.
   - Trigger-B: TCF notifies the status only of UEs relevant to TCF anytime when the RAN congestion status changes in an eNB.

   Since Trigger-A needs to request the RAN congestion status of all UEs registered in CDB,

 ISBN: 978-1-61208-398-8

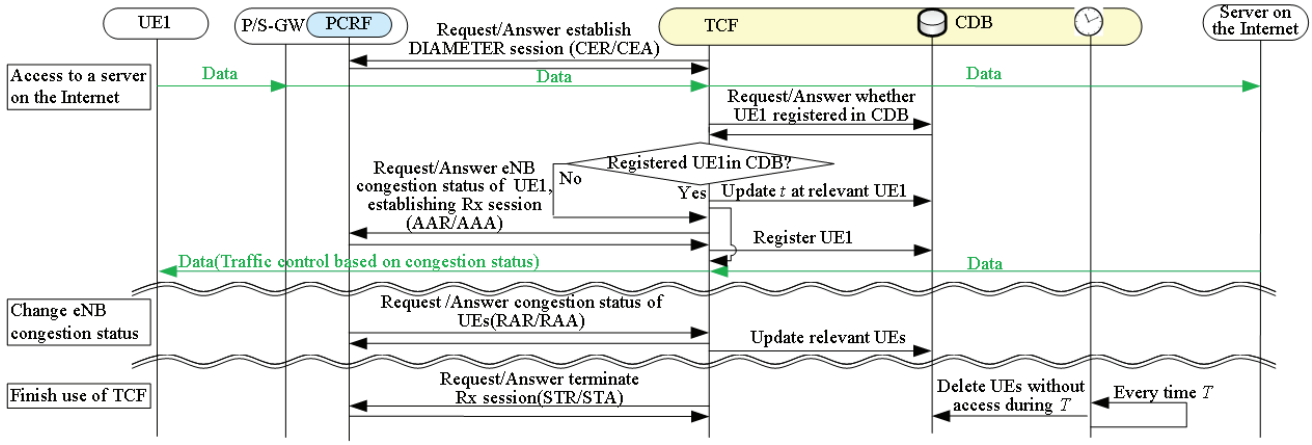


Figure 2. Sequence flow of TCF.

we select Trigger-B, by which a lower volume of RAN congestion status notifications are expected.

5. In the case that the IP address of a certain UE is changed or reallocated to another UE, the CDB ends up storing the invalid RAN congestion status. Therefore, UE data without access for a certain period *T* are deleted from CDB.
6. Whenever the RAN congestion status of UE in CDB is accessed, the corresponding access time *t* on the CDB is updated.

B. Congestion Status Database (CDB)

Table I shows an example of CDB at each TCF. Each record in CDB consists of the UE's IP address, Rx session ID to identify TCF, RAN congestion level *C* at the eNB that UE is now visiting, access time *t*, Cell ID that identifies the eNB, International Mobile Subscriber Identity (IMSI), and Mobile Station International Subscriber Directory Number (MSISDN). When reading a record, the UE's IP address is used as a search key. The values of Cell ID, IMSI, and MSISDN are set based on the obtained information from PCRF.

TABLE I. EXAMPLE OF CONGESTION STATUS DATABASE (AT TCF1)

* CDB key

IP Address *	Rx Session ID	Congestion Level *C*	Access Time *t*	Cell ID	IMSI	MSISDN
IPa	TCF1+IPa	3	*t*1	1	00	aaa
IPb	TCF1+IPb	1	*t*2	2	11	bbb
IPc	TCF1+IPc	2	*t*3	1	22	ccc
:	:	:	:	:	:	:

The CDB accesses occur at three timings (and actions) as shown below.

- *E1*: when a UE accesses TCF (read record and update *t*, or, insert new record)
- *E2*: when a TCF receives a RAN congestion status change at an eNB from PCRF (update *t*)
- *E3*: when a certain period *T* passes from the last access time (delete record)

Note that, at *E2* and *E3* timings, multiple CDB accesses will occur at the same time in proportion to the number of corresponding UEs.

C. Sequence Flow of TCF

The sequence flow of TCF is as follows (see Fig. 2);

1) After a TCP connection between TCF and PCRF is established, a DIAMETER session is established to exchange its identification and functional information (such as protocol version, supported DIAMETER application, and security mechanism) by Capabilities-Exchange-Request/Answer DIAMETER messages (CER/CEA).
2) When a UE accesses a server on the Internet via a TCF, the TCF checks whether the IP address of the UE has already been registered in its CDB. If yes, the TCF updates access time *t* and goes to 5).
3) If no, the TCF sends the UE information (IP address and Rx session ID) to PCRF. The PCRF replies with the congestion status of the eNB that the UE is visiting by AA-Request/Answer Rx messages (AAR/AAA). Then, the Rx session for the UE is established between TCF and PCRF.
4) The TCF registers the UE information, received the congestion level *C*, and access time *t* at CDB.
5) The TCF controls the traffic to the UE based on the congestion level *C*.
6) When the PCRF detects the congestion level changes at the eNB, it sends the corresponding information of the UE and its congestion level to the TCF by Re-Auth-Request/Answer Rx messages (RAR/RAA), and the TCF updates the congestion level *C* of the UE and access time *t*.
7) The TCF finds records in each of which time *T* has been passed from the last access at a constant period *T*, then deletes the matched records from its CDB. In addition, the TCF and the PCRF exchange the termination command by Session-Termination-Request/Answer Rx messages (STR/STA) in order to terminate the relevant Rx session.

 ISBN: 978-1-61208-398-8

IV. EXPERIMENTAL EVALUATION

In order to validate the effectiveness and feasibility of our proposal, we set up an experimental traffic control system on a physical PC (HP DL380p Gen8, 16 core (Intel Xeon E5-26600@2.20GHz), 128GB memory, 400GB HDD, VMware ESXi 5.1), where a commercial vEPC software is running as S/P-GW+PCRF [9] and two types of our developed software are running as TCF1 and TCF2, respectively. We arranged three Virtual Machines (VMs) for S/P-GW+PCRF, TCF1, and TCF2 as shown in Fig. 3. The resources allocated to both TCFs are equivalent, but OSes are different as shown in Table II. The CDB on each TCF is implemented with SQLite3 [13]. The experimental network has two cells. Each cell holds two UEs attached to two types of networks; emulated LTE and Wi-Fi.

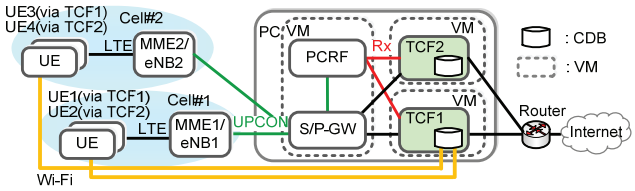


Figure 3. Experimental setup.

TABLE II. VM CONFIGURATION OF EACH TCF

Parameter	TCF1	TCF2
CPU	4vCPU	4vCPU
Memory	28GByte	28GByte
HDD	68GByte	68GByte
OS (Kernel)	Debian 7.1 (3.2.0-4-amd64)	CentOS 6.4 (2.6.32-358.el6.x86_64)

A summary of the traffic control functions is as follows; TCF1 aggregates two types of networks between UE and TCF1, i.e., LTE and Wi-Fi, to obtain higher bandwidth and/or offload the traffic [14]. TCF1 controls the traffic for LTE (it is the target network) based on RAN congestion status *C*. Meanwhile, TCF2 provides a proxy function to aggressively pre-fetch a web content that the UE is expected to access next [15] only when (LTE) RAN is not congested (pre-fetching ON). When (LTE) RAN is congested, such a pre-fetching is not applied (pre-fetching OFF). Note that, we verified the pre-fetching ON/OFF whether the pre-fetch tags are included or not at the header in the downloaded web content.

We set the routing in S/P-GW so that UE1 and UE3 can access the Internet via TCF1, and so that UE2 and UE4 can do it via TCF2.

A. Verification of TCF Behavior based on Congestion Status

First, we verify that the implemented TCFs can control the traffic based on the congestion status *C*. In order to confirm it, each UE accesses the web server on the Internet via the corresponding TCF while the congestion level *C* is varied from 1 to 3 (3 is the most congested) as shown in Table III. In this experiment, UE1 and UE3 via TCF1 start downloading a big size file (763MB) from the web server, UE2 and UE4 via TCF2 access the web server 5 times each in total while *C* changes 1→2→3→1 at eNB1 every 60 seconds.

TABLE III. TCF PARAMETERS AND ACCESS SCENARIOS

TCF #	Function	Congestion level *C* and function ON/OFF	Download files and access timing
1	Control LTE bandwidth	*C*=3 : ON (0.4Mbps) *C*=1, 2: OFF (no control)	Download a file (763MB) from a web server while changing *C*.
2	Pre-fetching	*C*=1: ON *C*=2, 3: OFF	Download Google web content 5 times each in total.

1) Per-UE and per-TCF based traffic control

Figure 4 shows the traffic control results of UE1 and UE2 measured at each TCF when *C* changes in eNB1. In Fig. 4 (a) and (b), the X-axis expresses the elapsed time [sec] and *C* was changed at the times indicated by red arrows.

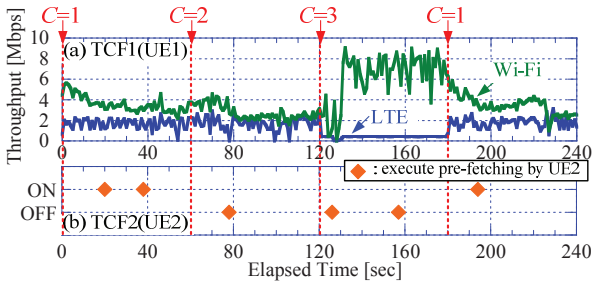


Figure 4. Example of traffic control.

From these plots, TCF can control the traffic based on *C*. As shown in Fig. 4 (a), with UE1 via TCF1, the throughput of LTE changes approximately 2.0→2.0→0.4→2.0 [Mbps]. The LTE bandwidth of UE1 is limited and traffic is offloaded toward Wi-Fi, i.e., traffic control is ON when *C* is 3, while that of UE2 changes to OFF when *C* is 2 and 3. This proves that TCF can control the traffic on a per-UE basis.

In addition, according to the result of UE2 via TCF2 as shown in Fig. 4 (b), the pre-fetching is ON when *C* is 1 and OFF when *C* is 2 and 3, while the throughput of TCF1 in Fig. 4 (a) is almost the same when *C* is varied from 1 to 2, irrespective of the ON/OFF change of TCF2. This proves that TCF1 and TCF2 can control the traffic independently.

2) Per-cell based traffic control

Table IV shows the change of the traffic control status at each UE in cell#1 and cell#2 while *C* changes in cell#1 only. When *C* is varied from 3 to 1, the traffic control statuses at UE1 and UE2 are changed, while those at UE3 and UE4 are kept the same in both ON/OFF cases. Note that there are two cases of OFF→OFF in cell #1 as indicated by red italic characters in Table IV. This reason is that TCF1 is OFF when *C* is 1 and 2, and TCF2 is OFF when *C* is 2 and 3 as shown in Table III. Hence, the TCFs in cell#1 can control the traffic based on *C* in these cases. This proves TCF can control the traffic based on a per-cell basis.

 ISBN: 978-1-61208-398-8

These results including 1) indicate that our proposed implementation design of the traffic control functions is feasible for controlling the traffic based on the RAN congestion status of corresponding UEs by each TCF.

TABLE IV. CHANGE OF THE TRAFFIC CONTORL STATUS AT EACH UE AND CELL (VARIED *C* AT CELL#1)

Cell#	eNB#	UE#	TCF#	*C*1→*C*2	*C*2→*C*3	*C*3→*C*1
1	1	1	1	*OFF → OFF*	OFF → ON	ON → OFF
		2	2	ON → OFF	*OFF → OFF*	OFF → ON
2	2	3	1	OFF → OFF	OFF → OFF	OFF → OFF
		4	2	ON → ON	ON → ON	ON → ON

: Changed traffic control status. *Italic* : Same traffic status in cell#1.

B. Performance Evaluation of CDB

As our proposed system accesses its CDB on a per-UE basis, it is important to perceive the processing performance of the CDB for the number of UEs. To evaluate it, the processing time of CDB at each TCF is measured in four cases; *read*, *insert*, *delete*, and *update*. Specifically, we measure the processing time for accessing *L* records 10 times in each case, where *L* denotes the registration number of UEs (= the number of records in CDB).

Figure 5 shows the processing time at TCF1 and TCF2 when the number of UEs is changed. The processing time is increased in the following the order, *read* < *insert* < *delete* < *update*.

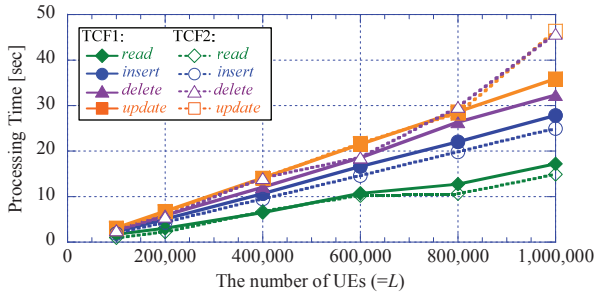


Figure 5. CDB processing time of UEs.

In the case of TCF1, the processing time is (almost) linearly increased as *L* increases. On the other hand, in the case of TCF2, it is also linearly increased when the number of UEs is smaller (*L*<=600,000), but it is rapidly increased when the number of UEs is larger. Specifically, when *L* is 1,000,000, *read* and *update* in TCF1 are 17.17 and 35.85 seconds, respectively, and those of TCF2 are 14.88 and 46.34 seconds, respectively.

According to [16], the system performance of the commercial PCRF is approximately 25,000 [TPS] per server. The processing times of *read* and *update* in TCF are 58,241 and 27,894 [TPS], respectively. This indicates that the performance of TCF is acceptable for practical use.

The *update* processing time is approximately from 2 to 2.5 times longer than *read*. Furthermore, the processing times in TCF1 is shorter than TCF2 at *L*<600,000, whereas it is reversed in *delete* and *update* at *L*>800,000. We assume that this is due to the kernel version because it is the only differential factor between TCF1 and TCF2. This result implies that CentOS is suitable for lower *L*, and that Debian is suitable for larger *L*.

These experimental results indicate that the performance of our proposed system is affected by the *update*, which requires the longest processing time. In addition, the OS (or kernel) selection seems important for achieving good performance at the CDB.

V. DISCUSSION

Based on the results in Section IV *B*, we discuss the applicability of UPCON to MNO. As one example, we refer to Japanese MNO data that indicates the penetration rate in the population of LTE is in the top 3 in the world. Table V shows the number of eNBs and subscribers (i.e., UEs) that MNOs announced in March 2014 [17][18] and the number of UEs per eNB in the Japanese top 3. We analyze how CDB load is carried in regard to the number of UEs and eNBs. We focus on *update* that requires the longest processing time. For simplicity, we assumed that each MNO has one EPC, the number of UEs per eNB is equal in each MNO, a uniform time interval is required to notify the congestion status, and all UEs use the same TCF.

TABLE V. JAPANESE MNOS' DATA

MNO (M_i)	1	2	3
#eNBs (N_i)	97,755	61,062	34,048
#UEs (H_i)	40,522,000	63,105,200	35,924,800
#UEs per eNB (K_i=H_i/N_i)	414.53	1033.46	1055.12

In Table V, M_i (i=1, 2, 3) denotes each MNO. N_i and H_i denote the number of eNBs and UEs at M_i, respectively. K_i indicates H_i per N_i. Let p_{ij} denote the *update* processing time of one million UEs at TCFj (j=1, 2), the per-eNB *update* processing time is given by $K_i \cdot p_{ij}$/1,000,000, which we denote as v_{ij}. From the result of Section IV *B*, v_{11}=0.015, v_{21}=0.037, and v_{31}=0.038 seconds at TCF1 and v_{12}=0.019, v_{22}=0.048, and v_{32}=0.049 seconds at TCF2, respectively.

Assuming that the congestion status is notified every r seconds from each eNB to PCRF (we call r "notification interval"), TCFj at M_i needs to finish the *update* process within u_i=r/N_i seconds on average, i.e., v_{ij}<u_i.

First, we analyze the effect of r. Figure 6 shows the relationships between r, u_i, and v_{ij}. For instance at M_1, u_1 is 0.037 seconds when r is 3,600 seconds. Here, TCF1 can process *update* because it satisfies v_{11}<u_1. In contrast in the case of r is 600, the *update* cannot be finished within the given time because of v_{11}>u_1. Let R_{ij} denote the threshold value of r by which TCFj at M_i could process *update*, we can easily obtain R_{11} as 1,452 seconds. In the case of TCF2, R_{12} is 1,877 seconds.

Furthermore, for instance at TCF1, in the case of M_2 and M_3, although u_3 is two times as much time as u_2 in order to finish the *update* process, v_{21} and v_{31} are almost the same value because v_{ij} depends on K_i. In addition, in the case of M_1 and M_3, although N_1 is three times larger than N_3, R_{11} and R_{31} are almost the same value because R_{ij} depends on H_i.

 ISBN: 978-1-61208-398-8

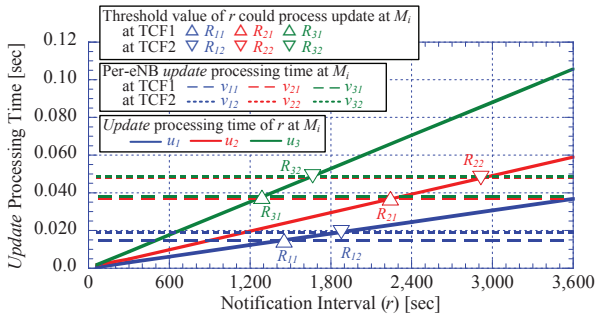


Figure 6. Notification interval of the RAN congestion status from eNB to PCRF vs. *update* processing time.

Next, when the congestion statuses are notified more frequently than R_{ij}, more processing resources are required for TCF, e.g., arranging multiple physical (and/or virtual) machines (we call them "units") for TCF. Hence, we analyze the relationships between r and the number of units for TCFj. Let s_{ij} denote the number of units for TCFj at M_i, and it is given by $s_{ij}=v_{ij}\cdot N_i/r$.

For instance, at M_1 in the case of TCF1, s_{11} is 24.4 units when r is 60 seconds. In contrast, in the case of r is 600 seconds, s_{11} is 2.44 units. Therefore, s_{ij} is an extremely large number when r is short. In the case of TCF2, s_{12} is 31.0 units when r is 60 seconds. In contrast in the case of r is 600 seconds, s_{12} is 3.10 units. Hence, the selection of r and s_{ij} is important.

As one example according to [16], many millions of subscribers are handled by a single PCRF server composed of many blades. Hence, arranging multiple physical machines for TCF seems a practical solution. We leave the effective configuration of multiple servers/blades as future work.

VI. CONCLUSION

In this paper, we propose an implementation design of the TCF working with vEPC. It controls the traffic in accordance with RAN congestion status, which is supposed to be notified by eNB to EPC and its behind systems based on UPCON. Our implementation design covers the executions of multiple TCFs independently based on the RAN congestion status, which can be retrieved from each TCF with a reasonable overhead. We implemented the proposed system with commercial vEPC software, which includes EPC functions, such as S/P-GW and PCRF. The experimental results show that the proposed system can control the traffic based on the RAN congestion status, and demonstrate the performance of CDB. Furthermore, we discuss the applicability of TCF to MNOs about the relationship with the notification time from eNB to PCRF and the number of TCFs. As future work, we will verify that achieving good performance at the CDB to select some OSes (or kernels) and the effective configuration of multiple servers/blades. In addition, we will evaluate the performance of the CDB by varying the number of eNBs and UEs.

REFERENCES

[1] Cisco Systems Inc, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

[2] D. Kutscher, H. Lundqvist, and F. G. Mir, "Congestion Exposure in Mobile Wireless Communications," Proc. GLOBECOM, Dec. 2010.

[3] A. Shadmand and M. Shikh-Bahaei, "TCP Dynamics and Adaptive MAC Retry-Limit Aware Link-Layer Adaptation over IEEE 802.11 WLAN," in Proceedings of CNSR 2009, pp.193-200, May 2009.

[4] A. Zolfaghari and H. Taheri, "Queue-Aware Scheduling and Congestion Control for LTE," Proc. ICON 2012, pp.131-136, Dec. 2012.

[5] 3GPP TR 36.819, "Coordinated multi-point operation for LTE physical layer aspects," http://www.3gpp.org/DynaReport/16819.htm.

[6] 3GPP TS 23.705, "System Enhancements for User Plane Congestion management," http://www.3gpp.org/DynaReport/23705.htm.

[7] M. Shehada, "Overview of 3GPP Study Item UPCON," http://www.ikr.uni-stuttgart.de/Content/itg/fg524/Meetings/2012-03-13-Muenchen/01_ITG524_Munich_Shehada.pdf.

[8] A, Maeder, S. Schmid, and Z. Yousaf, "Towards User-plane Congestion Management in LTE EPS," http://www.slideshare.net/zahidtg/towards-userplane-congestion-management-in-lte-eps.

[9] ANMCC, http://affirmednetworks.com/products/mobileContentCloud.php.

[10] 3GPP TS 23.203, "Technical Specification Group services and System Aspects; Policy and charging control architecture," http://www.3gpp.org/DynaReport/23203.htm.

[11] 3GPP TS 29.214 "Technical Specification Group Core Network and Terminals; Policy and Charging Control over Rx Interface Point (Release11)," http://www.3gpp.org/DynaReport/29214.htm.

[12] P. Calhoun, J. Loughney, E. Guttman, G. Zorn, and J. Arkko, "Diameter Base Protocol," IETF, Informational RFC 3588, Sep. 2003, http://tools.ietf.org/html/rfc3588.

[13] SQLite, http://www.sqlite.org/download.html.

[14] A. Tachibana, T. Yoshida, M. Shibuya, and T. Hasegawa, "Implementation of a Proxy-based CMT-SCTP Scheme for Android Smartphones," Proc. IEEE WiMob 2014, pp.664-669, Oct. 2014.

[15] T. Goto, A. Tachibana, and T. Hasegawa, "A Study on Prefetching Control Method for Web Traffic based on Congestion Information of Base Stations," Proc. IEICE Spring Conference, B-11-14, Sep. 2013.

[16] Broadband Traffic Managment, "PCRF Performance: Opnet vx. Comverse Vs. Others," http://broabandtrafficmanagement.blogspot.jp/2013/01/pcrf-performance-openet-vs-comverse-vs.html.

[17] Ministry of Internal Affairs and Communications, "Information & Communications Statistics Database," http://www.tele.soumu.go.jp/j/musen/toukei/index.htm (in Japanese).

[18] Telecommunications Carriers Association, "Number of Subscribers," http://www.tca.or.jp/english/index.html.

 ISBN: 978-1-61208-398-8

A Novel Protocol for Interference Mitigation in MIMO Femto Cell Environment

Zuhaib Ashfaq Khan

Department of Electrical Engineering
COMSATS Institute of Information Technology
Attock, Pakistan
Email: zakpassion@gmail.com

Muhammad Hasanain Chaudary
and Juinn-Horng Deng

Department of Communications Engineering
Yuan Ze University, Taiwan
Email: chaudaryh, jhdeng.hank@gmail.com

***Abstract*—In this paper, a novel protocol to mitigate co-channel interference issue acquiring higher diversity gain is introduced in Multiple Input Multiple Output (MIMO) femto cell environment. The scheme works under three time slots using Alamouti Space-time block code (STBC) and MIMO gains to achieve full diversity order of four. A network based on wireless access consists of two femto users points, and destination terminals as femto base stations each equipped with two antennas respectively. The performance is analyzed using 16-Quadrature Amplitude Modulation (QAM) modulation scheme. The results are investigated over independent and identical ($i.i.d$) fading channel. Subsequently, the obtained results by simulation show that the proposed novel transmission protocol provides interference mitigated and better signal quality of uplink femto users in cooperative wireless networks.**

Keywords–Femto Cooperative (Fe-COPE); Interference mitigation; MIMO-Femto environment; Alamouti Coding.

I. INTRODUCTION

People all over the globe are stepping ahead to the revolution of wireless broadband services in the form of 5G networks [1]. As fifth generation networks deal with the ultra high frequency regions of the spectrum, hence they consequently characterize by shorter wavelengths and the original signal tends to dissipate even after traveling up to tens of meters due to several penetration losses that leads to a situation of retaining smaller cell size. Furthermore, many shadowed and non-coverage regions could be found under the macrocells coverage areas such as at the edge of the cell, thick walls, indoor environment and basements, etc. It is a big challenge for the network operators to find an efficient solution to deal with this increasing data traffic demand and satisfy their customers by providing high quality services.

Femtocell deployment is a hot research stream for the researchers striving to improve the quality of service within a macrocell. These are low power tiny base stations installed within an office or a residential area to improve cellular coverage within a building. Commercially, it is sold with different brand names like Airave (Sprint), Microcells (AT&T) and Network Extender (Verizon) in the market [2]. The connectivity of smart phones, cellular phones and portable devices is increased through femtocells for cellular networks especially in those areas where coverage of large cells is weak or intermittent. There are many potential advantages of femtocells for both users and network operators. It provides better coverage and many additional services to the users and enables them to add the enhanced experience and extra capabilities to the existing broadband services. However, despite of all these characteristics, the main drawbacks include the higher interference issues and increased in the number of handovers that results in faster battery discharge. Network operators can generate more revenue by providing many additional services and better coverage to users up to their contentment. Therefore, the demand of femtocell has been increased since last decade and many commercial mobile operators have shown their keen interest to expand this technology through 3G to 4G and LTE/5G.

Multiple Input Multiple Output (MIMO) systems are widely considered to provide greater spectral efficiency and enhanced capacity. The transmission rate is badly affected due to the several limitations on the modulation scheme with a limited frequency band despite of offering guaranteed high Signal-to-Noise Ratio (SNR) in short range communications. MIMO systems have drawn researchers' [3][4] attention because of the fact that they can increase the data rate without even expanded frequency band by the use of multiple antennas over the whole transceiver design. The proposed protocol exploits this advantage of MIMO by achieving twice the gain achieved in conventional femtocells to mitigate the interference occurred in the overlapping areas of adjacent cells.

As femtocells are in microcells, there is a greater probability of incurring interference between two adjacent femtocells, especially in the overlapping areas of consecutive cells. Interference severely degrades the quality of service affecting the actual gain of the transmitted signal. Researchers have put serious efforts on the board to mitigate this interference in last couple of years using various schemes [5][6], but all these techniques cover different aspects and have not considered MIMO in three time slots for the minimal interference in femtocells [7][8].

This work analyzes the effect of interference on overall antenna gain in overlapping areas of two adjacent femtocells in MIMO environment and proposes a protocol for mitigating this interference in MIMO femto environment using three time slots. Antenna gain is finally achieved four times given by equations in the overall system model by the exploitation of MIMO. Simulation results using MATLAB clearly argue that the proposed protocol is an efficient technique for mitigating the interference in MIMO femto environment using three time slots as compared to many others.

To the best of our knowledge, no one has yet investigated MIMO femto cells scenario with Alamouti coding technique and exploits the cooperative scenario to mitigate interference issue in femto cell environment. The remainder of the pa-

 ISBN: 978-1-61208-398-8

per is organized as follows: Section 2 presents the model of the system, transmission scheme and the channel of the model considered. Sections 3 and 4 discusses in detail the input-output and closed-form expressions of a system at the respective base stations and then will derive the expression for SNR for both users respectively. Section 5 depicts the closed-form expressions for Nakagami-m fading channel using error probability analysis of a system. Section 6 discusses the simulation results obtained in terms of a comparison with analytical and Matlab obtained curves. Finally, Section 7 draws the conclusions of the prosposed scheme.

II. System Model

In this section, a novel approach is used to mitigate co-channel interference in MIMO scenario in order to exploit much higher diversity gain. Multiple antennas are assumed to be equipped at the user's terminals. For our system in consideration, we assume two (2) number of antennas equipped at the transmitter and receiver end.

For simplicity, MIMO based Fe-COPE system is furnished with two nodes at transmitter and receiver end. Both nodes are equipped with two antennas to make 2x2 MIMO system and exploit the gain after mitigating the interference from interferer node using three time slots. The communication is done in such a way that user nodes act as a relay in the third slot. The femto user 1 (node 1) has antennas S_1 and S_2 whereas, femto user 2 (node 2) has antennas S_3 and S_4 respectively. Both nodes will work as a relay and will follow Alamouti coding in the final slot of transmission to exploit MIMO and Alamouti gain. The transmission protocol is explained in the Table I.

The transmission takes place in three separate slots between source and end terminals. It is also assumed that the energy is normalized and at the respective source terminals, the power of the signal is normalized to unity with E = $\{ |S_i^2| \}$. Assuming equal noise variance N_o for the additive white Gaussian Noise (AWGN). The perfect channel condition is assumed at destination ends. The mathematical expressions for the input-output (I/O) equations, SNR relationship, & moment generating function (MGF) expressions will be discussed in the next section for different fading environments.

A. Cooperative Transmission and Channel Model

In this section, the equivalent expression for end-to-end transmission of SNR for both users at femto base terminals is derived. The I/O relationship is discussed. MRC is used at receiver to accumulate and add-up the desired signals and with utilizing SNR equations, MGF expressions for both users are derived. The transmission can be clearly explained with the help of the system model picture. The detail schematics picture of channel distribution associated with each time slot is shown in Figure 1.

The figure clearly depicts that in the first time slot, both the first antennas of both nodes will transmit whereas, in the second time slot, the respective second antennas of both mobile stations (MS-1 and MS-2) will transmit. In the third slot, both antennas of the mobile stations will transmit using Alamouti scheme approach to the respective femto base stations (BS-1 and BS-2), respectively.

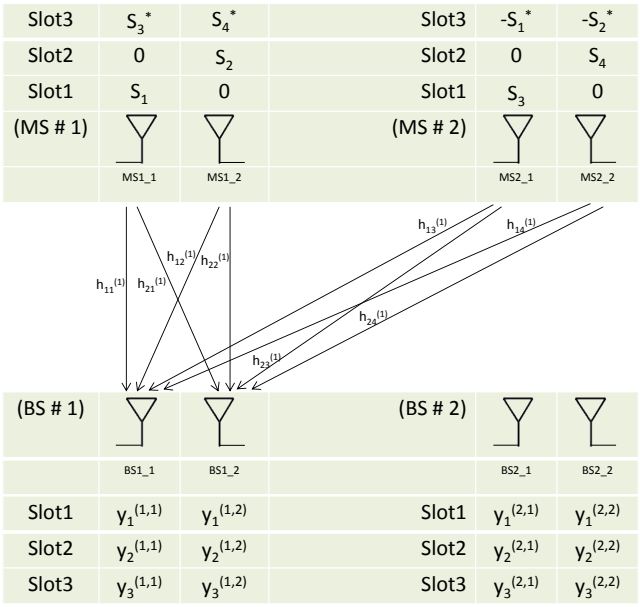


Figure 1. Schematic picture of channel distribution associated with each time slot

III. I/O and Closed-form Expressions

The mathematical expressions for an equivalent SNR relationship and respective closed-form expression for both femto stations are derived in this section, as follows.

A. I/O Expression

The received signals for both antennas, i.e., ANT1 and ANT2, etc., at both femto base stations (BS-1) in three different slots from respective two femto users is mathematically elaborated below.

Considering receptor BS1:

I/O for ANT1:

$$y_1^{(1,1)} = h_{11}^{(1)} S_1 + h_{13}^{(1)} S_3 + n_1^{(1,1)} \quad (1)$$

$$y_2^{(1,1)} = h_{12}^{(1)} S_2 + h_{14}^{(1)} S_4 + n_2^{(1,1)} \quad (2)$$

$$y_3^{(1,1)} = h_{11}^{(1)} S_3^* + h_{12}^{(1)} S_4^* - h_{13}^{(1)} S_1^* - h_{14}^{(1)} S_2^* + n_3^{(1,1)} \quad (3)$$

I/O for ANT2:

$$y_1^{(1,2)} = h_{21}^{(1)} S_1 + h_{23}^{(1)} S_3 + n_1^{(1,2)} \quad (4)$$

$$y_2^{(1,2)} = h_{22}^{(1)} S_2 + h_{24}^{(1)} S_4 + n_2^{(1,2)} \quad (5)$$

$$y_3^{(1,2)} = h_{21}^{(1)} S_3^* + h_{22}^{(1)} S_4^* - h_{23}^{(1)} S_1^* - h_{24}^{(1)} S_2^* + n_3^{(1,2)} \quad (6)$$

whereas, the received signals $y_k^{(i,j)}$ is at k-th time slots from both users, where ($k \in 1, 2, 3$), i denotes the respective base station 1, and j denotes the number of receiving antenna ($j \in 1, 2$). The noises $n_k^{(i,j)} \in \mathcal{N}(0, N_o)$, following AWGN, where

TABLE I. MIMO FEMTO PROTOCOL

Time Slot/Source Node	S_1	S_2	S_3	S_4	R_1	R_2	R_3	R_4
1	$\surd$		$\surd$					
2		$\surd$		$\surd$				
3					$\surd$ (S_3^*)	$\surd$ (S_4^*)	$\surd$ $(-S_1^*)$	$\surd$ $(-S_2^*)$

$(k \in 1, 2, 3)$ time slots respectively. Without loss of generality, assuming the level of the energy of the signal normalized to unity. The channel co-efficients undergoes the Nakagami-m fading [9] environments respectively. Rayleigh fading [3][6][9] is experienced as a special case when for Rician [6], K goes to 0, and for Nakagami, m goes to 1.

1) Algorithm: **Step 1: MIMO Detector (ML or SML Methods)**

Now, if $Norm(H_1^{(1)}) > Norm(H_2^{(1)})$, the detection process includes:

$$\begin{bmatrix} y_1^{(1,1)} \\ y_1^{(1,2)} \end{bmatrix} = \begin{bmatrix} h_{11}^{(1)} & h_{13}^{(1)} \\ h_{21}^{(1)} & h_{23}^{(1)} \end{bmatrix} \begin{bmatrix} S_1 \\ S_3 \end{bmatrix} + \begin{bmatrix} n_1^{(1,1)} \\ n_1^{(1,2)} \end{bmatrix} \quad (7)$$

whereas, $H_1^{(1)} = \begin{bmatrix} h_{11}^{(1)} & h_{13}^{(1)} \\ h_{21}^{(1)} & h_{23}^{(1)} \end{bmatrix}$. Using ML method, $\widehat{S}_1$, and $\widehat{S}_3$ will be detected.

else

$$\begin{bmatrix} y_2^{(1,1)} \\ y_2^{(1,2)} \end{bmatrix} = \begin{bmatrix} h_{12}^{(1)} & h_{14}^{(1)} \\ h_{22}^{(1)} & h_{24}^{(1)} \end{bmatrix} \begin{bmatrix} S_2 \\ S_4 \end{bmatrix} + \begin{bmatrix} n_2^{(1,1)} \\ n_2^{(1,2)} \end{bmatrix} \quad (8)$$

whereas, $H_2^{(1)} = \begin{bmatrix} h_{12}^{(1)} & h_{14}^{(1)} \\ h_{22}^{(1)} & h_{24}^{(1)} \end{bmatrix}$. Using ML method, $\widehat{S}_2$, and $\widehat{S}_4$ will be detected.

Step 2: Decision Feedback Cancellation

If $Norm(H_1^{(1)}) > Norm(H_2^{(1)})$, the algorithm is processed as,

$$\tilde{y}_3^{(1,1)} = y_3^{(1,1)} + h_{13}^{(1)}\widehat{S}_1^* - h_{11}^{(1)}\widehat{S}_3^* = -h_{14}^{(1)}S_2^* + h_{12}^{(1)}S_4^* + \tilde{n}_3^{(1,1)} \quad (9)$$

$$\therefore \begin{bmatrix} y_2^{(1,1)} \\ \tilde{y}_3^{(1,1)^*} \end{bmatrix} = \begin{bmatrix} h_{12}^{(1)} & h_{14}^{(1)} \\ -h_{14}^{(1)*} & h_{12}^{(1)^*} \end{bmatrix} \begin{bmatrix} S_2 \\ S_4 \end{bmatrix} + \begin{bmatrix} n_2^{(1,1)} \\ \tilde{n}_3^{(1,1)^*} \end{bmatrix} \quad (10)$$

$$\tilde{y}_3^{(1,2)} = y_3^{(1,2)} + h_{23}^{(1)}\widehat{S}_1^* - h_{21}^{(1)}\widehat{S}_3^* = -h_{24}^{(1)}S_2^* + h_{22}^{(1)}S_4^* + \tilde{n}_3^{(1,2)} \quad (11)$$

$$\begin{bmatrix} y_2^{(1,2)} \\ \tilde{y}_3^{(1,2)^*} \end{bmatrix} = \begin{bmatrix} h_{22}^{(1)} & h_{24}^{(1)} \\ -h_{24}^{(1)*} & h_{22}^{(1)^*} \end{bmatrix} \begin{bmatrix} S_2 \\ S_4 \end{bmatrix} + \begin{bmatrix} n_2^{(1,2)} \\ \tilde{n}_3^{(1,2)^*} \end{bmatrix} \quad (12)$$

Using Joint STBC Decoder, it gives the four (4) times diversity gain $\widehat{S}_2$ and $\widehat{S}_4$ are the detected signals with good diversity gains.

else

Similar to above equations in step number 1 and 2, using joint STBC decoder, $\widehat{S}_1$ and $\widehat{S}_3$ can be detected (4 diversity gain).

end

Step 3: Iterative cancellation to enhance detection performance

If $Norm(H_1^{(1)}) > Norm(H_2^{(1)})$, the algorithm is processed as,

$$\bar{y}_3^{(1,1)} = y_3^{(1,1)} + h_{14}^{(1)}\widehat{S}_2^* - h_{12}^{(1)}\widehat{S}_4^* = -h_{13}^{(1)}S_1^* + h_{11}^{(1)}S_3^* + \bar{n}_3^{(1,1)} \quad (13)$$

$$\therefore \begin{bmatrix} y_1^{(1,1)} \\ \bar{y}_3^{(1,1)^*} \end{bmatrix} = \begin{bmatrix} h_{11}^{(1)} & h_{13}^{(1)} \\ -h_{13}^{(1)*} & h_{11}^{(1)^*} \end{bmatrix} \begin{bmatrix} S_1 \\ S_3 \end{bmatrix} + \begin{bmatrix} n_1^{(1,1)} \\ \bar{n}_3^{(1,1)^*} \end{bmatrix} \quad (14)$$

$$\bar{y}_3^{(1,2)} = y_3^{(1,2)} + h_{24}^{(1)}\widehat{S}_2^* - h_{22}^{(1)}\widehat{S}_4^* = -h_{23}^{(1)}S_1^* + h_{21}^{(1)}S_3^* + \bar{n}_3^{(1,2)} \quad (15)$$

$$\therefore \begin{bmatrix} y_1^{(1,2)} \\ \bar{y}_3^{(1,2)^*} \end{bmatrix} = \begin{bmatrix} h_{21}^{(1)} & h_{23}^{(1)} \\ -h_{23}^{(1)*} & h_{21}^{(1)^*} \end{bmatrix} \begin{bmatrix} S_1 \\ S_3 \end{bmatrix} + \begin{bmatrix} n_1^{(1,2)} \\ \bar{n}_3^{(1,2)^*} \end{bmatrix} \quad (16)$$

Using Joint STBC Decoder, gives the four (4) times diversity gain $\widehat{S}_1$ and $\widehat{S}_3$ are the detected signals with good diversity gains.

else

Similar to above equations in step number 1 and 2, using joint STBC decoder, $\widehat{S}_2$ and $\widehat{S}_4$ can be detected (4 diversity

 ISBN: 978-1-61208-398-8

gain).

end

Step 4: Final step
In order to enhance the detection process, i.e., receive the signal after multiple trials resulting in higher performance gain, the output can be feedbacked to step number 2, and then go ahead with step number 3. This feedback process ensures the better detection probability.

IV. Signal to Noise Ratio Expression

The approach is followed by the algorithm steps and this section will conclude the SNR expression for user one at respective base station $BS-1$.

Using the final derived equations and summing up the received signals after detection, following steps can be used to derive SNR expression.

$$\begin{aligned}\tilde{y}_1^{(1,1)} &= \begin{bmatrix} h_{11}^{(1)^*} & -h_{13}^{(1)} \end{bmatrix} \begin{bmatrix} y_1^{(1,1)} \\ \bar{y}_3^{(1,1)^*} \end{bmatrix} \\ &= \left(|h_{11}^{(1)}|^2 + |h_{13}^{(1)}|^2\right) S_1 + \left(h_{11}^{(1)^*} n_1^{(1,1)} - h_{13}^{(1)} n_3^{(1,1)}\right)\end{aligned} \tag{17}$$

Similarly, for the second antenna, the received signal can be formulated as,

$$\begin{aligned}\tilde{y}_1^{(1,2)} &= \begin{bmatrix} h_{21}^{(1)^*} & -h_{23}^{(1)} \end{bmatrix} \begin{bmatrix} y_1^{(1,2)} \\ \bar{y}_3^{(1,2)^*} \end{bmatrix} \\ &= \left(|h_{21}^{(1)}|^2 + |h_{23}^{(1)}|^2\right) S_1 + \left(h_{21}^{(1)^*} n_1^{(1,2)} - h_{23}^{(1)} n_3^{(1,2)}\right)\end{aligned} \tag{18}$$

Accumulating the both antenna received signals at respective base station (BS-1) can be derived as follows,

$$\begin{aligned}\therefore \tilde{y}_1 &= \tilde{y}_1^{(1,1)} + \tilde{y}_1^{(1,2)} \\ &= \left(|h_{11}^{(1)}|^2 + (|h_{13}^{(1)}|^2 + |h_{21}^{(1)}|^2)|h_{23}^{(1)}|^2\right) S_1 + \\ &\left(h_{11}^{(1)^*} n_1^{(1,1)} - h_{13}^{(1)} n_3^{(1,1)} + h_{21}^{(1)^*} n_1^{(1,2)} - h_{23}^{(1)} n_3^{(1,2)}\right)\end{aligned} \tag{19}$$

Now, calculating the received signal power as follows,

$$Signal\ power = \left(|h_{11}^{(1)}|^2 + (|h_{13}^{(1)}|^2 + |h_{21}^{(1)}|^2) + |h_{23}^{(1)}|^2\right)^2 S_1 \tag{20}$$

Taking into account of the noise terms. The expectation of AWGN with zero mean i.i.d. noise, the noise power can be expressed as follows,

$$\begin{aligned}Noise\ power = E[(&h_{11}^{(1)^*} n_1^{(1,1)} - h_{13}^{(1)} n_3^{(1,1)} + h_{21}^{(1)^*} n_1^{(1,2)} \\ &- h_{23}^{(1)} n_3^{(1,2)})^2]\end{aligned} \tag{21}$$

$$NP = \left(|h_{11}^{(1)}|^2 + (|h_{13}^{(1)}|^2 + |h_{21}^{(1)}|^2) + |h_{23}^{(1)}|^2\right) \mathcal{N} \tag{22}$$

Hence, the final output SNR expression can be presented as follows,

$$Output\ SNR = \left(|h_{11}^{(1)}|^2 + (|h_{13}^{(1)}|^2 + |h_{21}^{(1)}|^2) + |h_{23}^{(1)}|^2\right) S/\mathcal{N} \tag{23}$$

This is the final expression of output SNR for user 1 at respective base station, BS-1. Similarly, using the similar approach, the output SNR for user 2 can be elaborated as follows,

$$Output\ SNR = \left(|h_{12}^{(1)}|^2 + (|h_{14}^{(1)}|^2 + |h_{22}^{(1)}|^2) + |h_{24}^{(1)}|^2\right) \frac{S}{\mathcal{N}} \tag{24}$$

After getting the desired SNR expression, the next step is to use the Q-function and MGF based approach to derive the close form expression for Bit Error Rate (BER) over Nakagami fading channels for both respective users.

V. Performance Analysis

A. Error Probability Analysis

In this section, the performance evaluation is done by taking into account the error probability (average SER) over Nakagami-m Fading Channel. The closely approximated Q-Function given in [9] as,

$$Q(x) \simeq \frac{1}{12} e^{-\frac{x^2}{2}} + \frac{1}{6} e^{-\frac{2x^2}{3}} \tag{25}$$

The probability of symbol error in the M-QAM system can be closely approximated as [9],

$$\overline{P}_{N_t \times N_r}(\gamma) \simeq \left(1 - \frac{1}{\sqrt{M}}\right) \left(\frac{1}{3} e^{-\frac{3\gamma}{2(M-1)}} + \frac{2}{3} e^{-\frac{2\gamma}{M-1}}\right) \tag{26}$$

The average SER can be tightly approximated by using close equation as,

$$\begin{aligned}P_{N_t \times N_r} &= \int_0^\infty \overline{P}_{N_t \times N_r}(\gamma) f_{N_t \times N_r}(\gamma) d\gamma \\ &\simeq \left(1 - \tfrac{1}{\sqrt{M}}\right) \left(\tfrac{1}{3} M_{N_t \times N_r}\left(\tfrac{3}{2(M-1)}\right)\right. \\ &\quad \left. + \tfrac{2}{3} M_{N_t \times N_r}\left(\tfrac{2}{M-1}\right)\right)\end{aligned} \tag{27}$$

Nakagami-m Fading Channel

To evaluate the average SER for our proposed system, we have adopted the MGF based approach. Thus we need to compute the MGF first for Nakagami-m fading channel by using the following expression,

$$\begin{aligned} M_{N_t\times N_r}(g)|_a &= \int_0^\infty e^{-g\overline{\gamma}(a+b+c+d)} f_{N_t\times N_r}(a)da \\ &= e^{-g\overline{\gamma}(b+c+d)} \int_0^\infty e^{-g\overline{\gamma}(a} \left(\frac{m}{\Omega}\right)^m \\ &\frac{1}{\Gamma(m)} a^{m-1} e^{-\frac{ma}{\Omega}} da \\ &= \left(\tfrac{m}{\Omega}\right)^m \tfrac{e^{-g\overline{\gamma}(b+c+d)}}{\Gamma(m)} \int_0^\infty a^{m-1} e^{-a\left(g\overline{\gamma}+\frac{m}{\Omega}\right)} da \end{aligned} \tag{28}$$

By using identity eq. 3.381.4 from [10], (28) become,

$$\begin{aligned} M_{N_t\times N_r}(g)|_a &= \left(\tfrac{m}{\Omega}\right)^m \tfrac{e^{-g\overline{\gamma}(b+c+d)}}{\Gamma(m)} \tfrac{1}{\left(g\overline{\gamma}+\frac{m}{\Omega}\right)^m}\Gamma(m) \\ &= \left(\tfrac{m}{\Omega}\right)^m \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-m} e^{-g\overline{\gamma}(b+c+d)} \end{aligned} \tag{29}$$

Also, by computing the MGF with respect to b, we get,

$$\begin{aligned} M_{N_t\times N_r}(g)|_b &= \int_0^\infty M_{N_t\times N_r}(g)|_a f_{N_t\times N_r}(b)db \\ &= \left(\tfrac{m}{\Omega}\right)^{2m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-m} \tfrac{e^{-g\overline{\gamma}(c+d)}}{\Gamma(m)} \\ &\int_0^\infty b^{m-1} e^{-b\left(g\overline{\gamma}+\frac{m}{\Omega}\right)} db \end{aligned} \tag{30}$$

By using identity eq. 3.381.4 from [10], (30) become,

$$\begin{aligned} M_{N_t\times N_r}(g)|_b &= \left(\tfrac{m}{\Omega}\right)^{2m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-m} \tfrac{e^{-g\overline{\gamma}(c+d)}}{\Gamma(m)} \tfrac{1}{\left(g\overline{\gamma}+\frac{m}{\Omega}\right)^m}\Gamma(m) \\ &= \left(\tfrac{m}{\Omega}\right)^{2m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-2m} e^{-g\overline{\gamma}(c+d)} \end{aligned} \tag{31}$$

By computing the MGF with respect to c, we get,

$$\begin{aligned} M_{N_t\times N_r}(g)|_c &= \int_0^\infty M_{N_t\times N_r}(g)|_b f_{N_t\times N_r}(c)dc \\ &= \left(\tfrac{m}{\Omega}\right)^{3m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-2m} \tfrac{e^{-g\overline{\gamma}(d)}}{\Gamma(m)} \\ &\int_0^\infty c^{m-1} e^{-c\left(g\overline{\gamma}+\frac{m}{\Omega}\right)} dc \end{aligned} \tag{32}$$

By using identity eq. 3.381.4 from [10], (32) become,

$$\begin{aligned} M_{N_t\times N_r}(g)|_c &= \left(\tfrac{m}{\Omega}\right)^{3m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-2m} \tfrac{e^{-g\overline{\gamma}(d)}}{\Gamma(m)} \tfrac{1}{\left(g\overline{\gamma}+\frac{m}{\Omega}\right)^m}\Gamma(m) \\ &= \left(\tfrac{m}{\Omega}\right)^{3m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-3m} e^{-g\overline{\gamma}(d)} \end{aligned} \tag{33}$$

Now, by computing the MGF with respect to d, we get,

$$\begin{aligned} M_{N_t\times N_r}(g)|_d &= \int_0^\infty M_{N_t\times N_r}(g)|_c f_{N_t\times N_r}(d)dd \\ &= \left(\tfrac{m}{\Omega}\right)^{4m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-3m} \tfrac{1}{\Gamma(m)} \\ &\int_0^\infty d^{m-1} e^{-d\left(g\overline{\gamma}+\frac{m}{\Omega}\right)} dd \end{aligned} \tag{34}$$

By using identity eq. 3.381.4 from [10], (34) become,

$$\begin{aligned} M_{N_t\times N_r}(g)|_d &= \left(\tfrac{m}{\Omega}\right)^{4m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-3m} \tfrac{1}{\Gamma(m)} \tfrac{1}{\left(g\overline{\gamma}+\frac{m}{\Omega}\right)^m}\Gamma(m) \\ &= \left(\tfrac{m}{\Omega}\right)^{4m} \left(g\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-4m} \end{aligned} \tag{35}$$

The average SER can be found after substituting the MGF expression in (27), as,

$$\begin{aligned} P_{N_t\times N_r} &\simeq \left(1-\tfrac{1}{\sqrt{M}}\right) \left(\tfrac{1}{3}\left(\tfrac{m}{\Omega}\right)^{4m} \left(\left(\tfrac{3}{2(M-1)}\right)\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-4m}\right) \\ &+ \left(\tfrac{2}{3}\left(\tfrac{m}{\Omega}\right)^{4m} \left(\left(\tfrac{2}{M-1}\right)\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-4m}\right) \\ &\simeq A_1 \left(\tfrac{1}{3}\left(\tfrac{m}{\Omega}\right)^{4m} \left(A_2\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-4m}\right) \\ &\left(+\tfrac{2}{3}\left(\tfrac{m}{\Omega}\right)^{4m} \left(A_3\overline{\gamma}+\tfrac{m}{\Omega}\right)^{-4m}\right) \end{aligned} \tag{36}$$

where $A_1 = \left(1-\frac{1}{\sqrt{M}}\right) = \left(1-\frac{1}{\sqrt{16}}\right) = 0.75$, $A_2 = \left(\frac{3}{2(M-1)}\right) = \left(\frac{3}{2(16-1)}\right) = 0.1$, and $A_3 = \left(\frac{2}{M-1}\right) = \left(\frac{2}{16-1}\right) = 0.1333$.

The above equation gives the theoretical average SER over Nakagami-m Fading Channel and in the results section, the theoretical average SER is plotted against the average SNR. The special case has been investigated also for Nakagami fading when m goes to one. So, that is the special case experienced as Rayleigh fading ($m = 1$).

VI. Simulation Results and Discussion

In this section, the performance of the MIMO based Fe-COPE protocol by plotting the analytical curves along with the simulated results is analyzed. The obtained theoretical results elaborate the performance evaluation of our proposed MIMO based Fe-COPE system for co-channel interference mitigation technique. The exact BER performance are analytically derived with the help of SER expression by using the MGF expressions under Nakagami-m and (Rayleigh fading as a special case) fading distributions respectively. The respective mathematical expressions in the form of theoretical results are drawn by using Mathematica 8 software and compared together with the obtained simulated results from Monte Carlo simulations using MATLAB over Rayleigh, and Nakagami fading environments, respectively. The conclusive obtained results for average performance parameter of BER are drawn w.r.t obtained SNR (E_b/N_o) in dB over 16-QAM constellation technique.

Figure 2 shows a good agreement between the analytical and simulated results over Nakagami fading channel. The figure presents the average BER analysis vs SNR in dB curve. It can be clearly seen that the performance improves in case of Fe-COPE proposed scheme due to gains of MIMO technique with diversity. Fe-COPE system helps in mitigating the co-channel interference in femto scenario but the interesting thing is that it exploits the channel gain of the interfered signal as well by using cell as a relay so this protocol is exploiting robust by taking into account the advantage of MIMO diversity gains and Alamouti gain, as well at both femto base stations. The results obtained showed that at value of BER = 10^{-2}, the performance of MIMO based Fe-COPE is better than normal Fe-COPE system due to four times diversity gain whereas in Fe-COPE the gains are of order two. The interesting part is that the BER curves go gradually better for the high SNR values as the value of m goes higher i.e. $m > 1, m = 2, 3$. The special case when $m = 1$ of Rayleigh fading is experienced and it can be verified with the theoretical and simulation results for MIMO based Fe-COPE protocol.

 ISBN: 978-1-61208-398-8

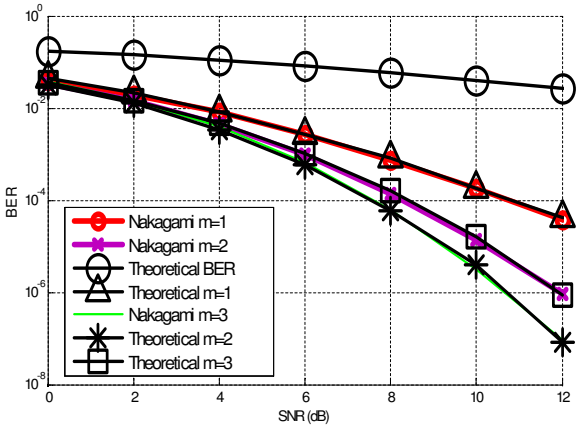


Figure 2. Simulation results for MIMO based Fe-COPE protocol over Nakagami Fading Channel

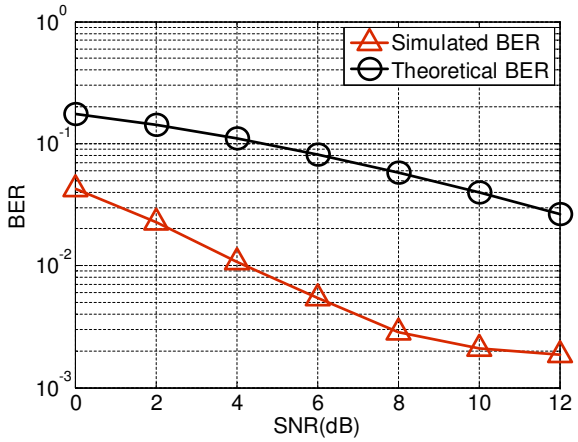


Figure 3. Simulation results of MIMO based Fe-COPE system in terms of BER Vs SNR curves for Rayleigh fading from S-D node over Rician fading environment

The results obtained for the Figure 3 are for the case when Rayleigh fading is considered from source nodes to destination in a Rician fading environment. It can be clearly seen that due to the MIMO and Alamouti gains, the results get better but for the high SNR values, the values reach to the saturation point and does not change much. It is a good solution for the low SNR values and depending on the different channel conditions, different approach can be adopted.

VII. Conclusion

A compact investigation on the performance is performed for MIMO-based femto cells over different fading channels. A novel protocol has been designed in order to mitigate interference issue in two users femto cell scenario by exploiting Alamouti coding gain as well as MIMO gains for 16-QAM modulation scheme. The closed form expressions are derived for the MIMO based Fe-COPE system over Rayleigh as special case, and Nakagami-m fading channels. The protocol involves five steps algorithm in order to reach full diversity (order of four) and then derive the I/O relationship to calculate SNR expression. Using the expression of SNR, closed form expression is calculated using MGF based approach for MIMO based Fe-COPE system. Later, the simulation results show the effectiveness of the protocol. The results are obtained using Mathematica and Matlab softwares in order to verify the results. The BER vs SNR curves are obtained for higher values of Nakagami factor m, it has been seen that for higher values there is a dramatic change in the results. The high SNR regime shows the promising results for the protocol and provide the maximum diversity order.

References

[1] Y. Kishiyama, A. Benjebbour, T. Nakamura, and H. Ishii, "Future steps of LTE-A: evolution toward integration of local area and wide area systems", IEEE Wireless Communications, February 2013, vol. 20, no. 1, pp. 12-18, doi: 10.1109/MWC.2013.6472194.

[2] L. Bao and S. Liao, "Scheduling heterogeneous wireless systems for efficient spectrum access", EURASIP Journal on Wireless Communications and Networking, Nov. 2010, vol. 11, pp. 1-14, doi: 10.1155/2010/736365.

[3] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", Wireless Personal Communications, March 1998, vol. 6, no. 3, pp. 311-335, doi: 10.1023/A:1008889222784.

[4] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays"', IEEE Signal Processing Magazine, Jan. 2013, vol. 30, no. 1, pp. 40-60, doi: 10.1109/MSP.2011.2178495.

[5] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks", IEEE Communications Magazine, April 2009, vol. 47, no. 4, pp. 74-81, doi: 10.1109/MCOM.2009.4907410.

[6] S. S. Arunachalam, S. Kishore Kumar, V. Manickam, and S. S. Murugan, "Performance analysis of spatial channel separation for interference mitigation in femtocellular systems", IEEE International Conference on Communications and Signal Processing (ICCSP), 4-5 April 2012, pp. 62-65.

[7] Y. Li, G. Zhu, and X. Du, "Aligning Guard Zones of Massive MIMO in Cognitive Femtocell Networks", IEEE Communication Letters, Feb. 2014, vol. 99, pp. 1-4, doi: 10.1109/LCOMM.2013.123113.131913.

[8] A. Chopra and B. L. Evans, "Joint Statistics of Radio Frequency Interference in Multiantenna Receivers", IEEE Transactions on Signal Processing, July 2012, vol. 60, no. 7, pp. 3588-3603, doi: 10.1109/TSP.2012.2192431.

[9] W. Kim, N. Kim, H. K. Chung, and H. Lee, "Performance Analysis and High-SNR Power Allocation for MIMO ZF Receivers with a Precoder in Transmit-Correlated Rayleigh Channels", IEEE Communications Letters, August 2012, vol. 16, no. 8, pp. 1304-1307, doi: 10.1109/LCOMM.2012.061912.120536.

[10] I. S. Gradshteyn and I. M. Ryzhik, "Tables of Integrals, Series and Products", Seventh ed., 84 Theobald's Road, London WCIX 8RR, UK, Elsevier Inc., 2007.

 ISBN: 978-1-61208-398-8

Improving Attack Mitigation with a Cost-sensitive and Adaptive Intrusion Response System

Rodion Iafarov, Ruediger Gad, Martin Kappes
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany
email: yafarovrs@gmail.com, {rgad, kappes}@fb2.fra-uas.de

***Abstract*—Because of the rise of the number of attacks in computer networks, mitigation measures have to be applied in an efficient manner. The time frame for attack mitigation is shortened what makes using classical manual intervention approaches less efficient. Even though the idea of Intrusion Response Systems (IRS) is not new, IRS are still not widely used. Potential users are typically afraid of inadequate reactions, which could worsen the situation or could even be used as a part of attacks. In this paper, we present a cost-sensitive, retroactive, adaptive, and preemptive IRS that is intended to support network administrators in the attack mitigation and decision making processes. Our approach aims on balancing the costs of responses and attacks, adapts to changing situations, and optimizes the selection of responses and response deployment locations. Experimental results obtained with an evaluation prototype show that our approach works and is feasible from a performance perspective.**

Keywords–Intrusion Response System; Risk Assessment; Impact Cost Assessment; Dynamic; Adaptive.

I. INTRODUCTION

The amount of attacks on Information and Communications Technology (ICT) increases, e.g., in 2013 an increase in the number of web-based targeted attacks of 25% and a 91% increase in targeted attacks campaigns could be observed [1]. Successful ongoing attacks may lead to severe consequences like significant monetary losses or may even endanger human health.

In order to avoid such consequences, it is paramount to mitigate attacks quickly and efficiently. Due to the increase of complexity and pace of attacks and intrusions, however, classical manual intervention is often not sufficient anymore. Weaknesses of classical manual intervention are the lack of speed, the requirement of expert knowledge, and the increasingly complicated response selection process.

Consequently, the necessity for more automated solutions has become obvious [2]. Intrusion Response Systems (IRS), which can be seen as an extension to Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) [3], have gotten more attention in recent years, particularly the combination of IRS with other approaches like decision-making [4][5].

The aim of automatic response approaches is to deal with attacks faster and more efficiently [6]. Fully automated systems, however, may trigger erroneous actions that may worsen the situation. As one consequence, system administrators usually perceive fully automated systems sceptical and are unwilling to hand over control to fully automated systems. Semi-automated approaches aim on solving this issue by allowing manual control while still accelerating the mitigation.

In this paper, we present an approach which improves the response selection process by supplying the user with pre-selected optimized response suggestions. Our solution takes advantages of existing methods and combines them for efficiently mitigating attacks. Furthermore, our approach allows additional human interaction and intervention, e. g., rolling back applied countermeasures or applying alternative actions, which may better fit in case the situation changes. With these mechanisms, we believe that the classical manual process can be significantly improved with respect to quality, speed, and deployment of reactions. While our system belongs to the class of manual response systems it can be extended to operate fully automated.

We assume that information about attacks is readily provided, e. g., by an IDS. The detection of attacks is beyond the scope of this paper.

In the following, we first present related work. In Section III, we introduce important requirements. Afterwards in Section IV we present our response selection approach. In Section V, we perform an assessment of our approach with a prototype. Finally in Section VI, we provide a conclusion of our findings and present an outlook on future work.

II. RELATED WORK

One line of research in the field of automatic and semi-automatic IRS deals with the development and application of cost models [7]–[9]. The objective of cost-sensitive approaches is to define a consistent metric, e. g, to balance the costs of attacks and responses or for decision-making. The classification proposed by Shameli-Sendi, Ezzati-jivan, Jabbarifar, and Dagenais [10] shows that recent researches pay more and more attention to the risk assessment mechanisms, adjustment and prediction abilities.

The approach to intrusion response proposed by Stakhanova, Strasburg, Basu, and Wong [6] introduces 4 types of costs: intrusion/response impacts on the system and intrusion/response operational costs. The response impact is evaluated based on the defined system goals and their importance, and the intrusion impact is evaluated with respect to the response ability to counter this damage. The response operational cost includes the costs for the setup of responses, the costs for the deployment of responses, and the costs of the data processing overhead needed to analyze the results of responses. The intrusion operational cost includes the baseline cost present for an attack and the actual damage that can be potentially caused by a successful attack. The main disadvantage of the model is absence of probabilistic analysis. Additionally, this model does not take into account combinations of responses that mitigate

 ISBN: 978-1-61208-398-8

attacks partially and sets of existing attacks and already applied reactions.

The cost model proposed by Lee, Miller, Stolfo, Fan, and Zadok [11] considers the potential harm of an attack and attack operational costs for monitoring and detection. In [11], Lee et al. do not split the response cost into categories but calculate an overall cost of acting against the attack. The main disadvantage of this model is uncertainty in the cost analysis due to incomplete or imprecise estimation and limitations related to the reconstruction of the model in case of metric changes.

The approach proposed in [9], defines the cost of damage caused by intrusion as the sum of intrusion impact on the system and cost of daily maintenance of various aspects of the detection system. The response cost is calculated cognate as the sum of impact on system and operation cost value.

In [7], Yaorui Wu and Shufen Liu use a cost model, which depends on probabilities referring to detection methods. Probabilistic techniques help to address risks of inadequate response deployment in case of detection errors.

Forecasting techniques can be applied to predict the possible development of attacks. Preemptive IRS, like the one presented in [6], use forecasting techniques to increase the accuracy of the countermeasure selection.

In [12], two types of response executions were defined: a burst model, which has no risk assessment mechanisms once the response has been applied and a retroactive model, which includes feedback mechanism that assesses the response effect based on the result of the applied response. In [13], an adaptive IRS was proposed that additionally introduces a response effectiveness index, which is used as quality indicator for responses.

Obviously, IRS should behave differently for different kinds of attacks. Therefore, attack classifications are required to allow an adequate response selection. The classification presented in [11], divides attacks into four main categories: probe, denial of service (DoS), remote to local (R2L), and user to root (U2R). In [14], Wu, Xiao, Xu, Peng, and Zhuang introduce another categorization for attacks similar to the one presented in [11]. In addition to the attack type, the categorization by Wu et al. also takes a location property into account, e. g., privilege escalating can be local or remote, resource depletion can be applied to host and/or network.

Risk and cost assessment are based on resource dependencies. The confidentiality, integrity, and availability (CIA) triad [15], e. g., can be used to define the importance of particular system resource security properties. Dependencies can be defined using the idea of a resource type hierarchy, as introduced in [16]. An additional graph-based data structure, the "system map", can be used to carry information about specific instances of the system. Within the system map, system resources are represented as vertices and dependencies between resources are defined as edges.

Existing models lack consistency and do not take advantages of each other. With our approach, we combine the aforementioned approaches in order to create a cost-sensitive, retroactive, adaptive, and proactive IRS with risk assessment based on resource dependencies, a dynamically evaluated cost model, and sustainable countermeasures according to the classification proposed in [10]. To the best of our knowledge, no such combination was presented before. Additionally, we consider the applicability of our solution in real networks in order to make a step towards using such solutions in real scenarios. Thus, we propose a semi-automated solution instead of a fully automated one, as it can be considered as more reliable. Our proposed solution is flexibly such that it can be adapted for being applied in varying environments and with varying sets of responses. Our model also aims on removing the lack of consistency with existing solutions by combining multiple different approaches.

III. System Resources Categorization

In our proposed approach, we use resources dependencies as a risk assessment criterion. As the first step of categorization, we assign importance values to the system resource security properties. As in [6], the importance of a particular system instance with respect to the CIA principles is defined by float values in the range $[0, 1]$, where 0 denotes minimum and 1 denotes maximum importance.

Additionally, as described in [17], the attack impact is split into three categories: none, partial, complete. The importance of the security properties defines how critical the loss of a certain attribute is. However, it is important to distinguish complete and partial affection to avoid unnecessary risk elevation, which may lead to inadequate response deployment. For each security property, we define two values, which correspond to partial and complete loss.

Dependencies between system resources are used for the risk assessment and impact cost calculation. During response selection, we have to take possible impacts on dependent system resources into account. Dependencies between resources can be declared as a directed graph.

To deal with cycles, the following procedure is used: At first we use a depth-first search and mark the states of the observed resources. When we observe a system resource for the second time, we check if there is an additional impact on the security properties in the new state. If yes, we assess the additional impact and go through the dependent nodes according to the new impact. Otherwise, we do not process dependent nodes as the previous assessment remains valid. This algorithm ends because maximum possible impact is defined, when all security properties are affected, and we always accumulate impact. When the maximum impact is reached for a resource node, the algorithm stops observing this node.

For defining dependencies, causal links are used. Conditions specify which security properties have to be affected to create a defined impact on the dependent system resource. For each dependency, the probability of the event is specified in order to define how probable the occurrence of the impact is. The defined probability is used to perform forensic analysis and to predict the possible attack development.

The dependency structure is also used to optimize the deployment location, which aims on minimizing risks and impact cost, as the cost for a response depends on the location where it is deployed or implemented. E. g., isolating an entire network affects all instances in the network whereas isolating a host only affects the host and the services on the host. Thus, in addition to finding a response with adequate costs, the deployment location can also be optimized.

The type attribute was added, because responses can have different impacts depending on the location of the resource they are applied on. The parameter corresponds to the affected instances type parameter of the response and is used to improve the accuracy of the response impact determination.

 ISBN: 978-1-61208-398-8

Additionally, the system resource physical location attribute has to be configured for the deployment procedure. Not all properties are involved in the response selection, as they were considered as less important due to the less impact on the process. Nevertheless, our proposed approach can be extended to consider additional attributes.

IV. Response Selection

The response selection procedure is performed as follows: At first the system resources and the impact of attacks are assessed. Then, a set of possible responses and the corresponding locations are determined. Afterwards, the impact of the selected responses is assessed. Finally, responses and locations are optimized.

The attack impact is defined by the target(s) and security properties, which it can affect. Based on the description of the environment, we can assess the possible impact on the target(s) and dependent system resources. For each response, besides impact, we define which security properties it can protect. Using this information we form subset of possible responses, which can mitigate an attack.

In difference to [2], our proposed algorithm also considers responses that do not mitigate attacks completely, but mitigate the impact on security properties, which are crucial for attacked system resource. The possibility to combine multiple responses for the attack mitigation is also taken into account. A response is added to the subset of possible responses if: the response completely mitigates an attack; or the response protects properties that are relevant for the system resource; or the response mitigates an attack partially while other responses exists that can protect the remaining relevant security properties.

Our proposed method forms the subset of the possible responses to select an optimized countermeasure aiming on minimizing overall risks and costs. A trade-off between attack and response impacts is performed and it is avoided to worsen the situation by wrong response deployment.

The location where a response is applied has to be determined as well. Response costs differ depending on the location and the costs can be minimized by optimizing the deployment location. The environment description is used to find all possible locations for the deployment. Then, the costs are assessed in order to determine the one, which provides minimal cost.

In order to optimize the response selection when multiple attacks are present in the system, our proposed solution takes new attacks, the set of current attacks, and the set of already applied responses into account for the calculations. With this approach it can be, e. g., identified if new attacks can be mitigated by already applied response or if it is possible to reduce costs by replacing previously deployed countermeasures Furthermore, if an attack was stopped, it is required to reconsider the deployed reactions and possibly apply a new set of responses, or cancel responses, which mitigated stopped attack.

A. Attack Cost

The attack cost is based on the impact on system resources and operational costs as defined in [6]. As we mentioned, it is assumed that the required information about an attack, including attack target, is provided by the detection mechanisms. We evaluate the probable attack development and consider system resources dependencies during cost assessment. The cost of an attack, denoted by a, is assessed by the function $atCost(a)$ and is calculated in accordance to 1.

$$atCost(a) = p_{det}(a)(\sum_{s \in S} p_{imp}(a)\omega(s) + opCost(a)), \quad (1)$$

where S is a set of security properties affected by the attack and $s \in S$ denotes a security property of a system resource. $\omega(s)$ is a function that computes the importance value of the affected security property s. $p_{det}(a)$ is a function that calculates the probability of the correct detection of an attack. $p_{imp}(a)$ is a function that computes the probability that an attack a will actually impact the system. Additional weights can be added as extension to the provided solution. The operational cost of attack a, denoted by function $opCost(a)$, is assigned by value in the range $[0, 1]$, as proposed in [6].

The probability of an attack impact is one of the required parameters. It allows to evaluate possible attack development. Additionally, probabilities of the impact on the attacked system resource dependencies are calculated to evaluate possibility of the impact and assess attack effect cost. The probability of the affection creates a non-increasing sequence, as for each next step, previous steps have to be successfully performed. The probability of the next step is calculated as multiplication of the probabilities of all required previous steps. This approach minimizes the risk of overestimating an attack and it improves the adequacy of the reaction. Additionally, we decrease the attack cost according to the detection confidence in order to avoid inadequate reactions in case of a detection error. This approach allows to minimize risks and perform additional investigation before actual deployment.

B. Response Cost

After the subset of possible responses is formed and the possible locations for the deployment were defined, the effects of each response in its possible deployment locations are assessed. The cost of a response, denoted by r, is calculated by the function $respCost(r)$. To avoid negative cost values, the base cost is initially set to the sum of all attacks persisting in the system, including the current attack, as shown in 2.

$$\sum_{a \in A} atCost(a), \quad (2)$$

where A denotes the set of all persisting attacks in the system. We also introduce an efficiency factor, denoted by the function $respEff(r)$. The efficiency property is specified for each response and is changed according to the results of response application. The efficiency is calculated as ratio of the number of successfully mitigated attacks by the response over the number of overall number of attacks we tried to mitigate by the response in accordance to 3.

$$respEff(r) = \frac{\#ofSuccessfullyMitigated}{overall\#ofTriesToMitigate}. \quad (3)$$

The response cost is decreased by the ability to mitigate the current attack and other persisting attacks influenced by the efficiency factor, as shown in 4:

$$respEff(r) \sum_{m \in M_r} atCost(m), \quad (4)$$

where M_r is the set of all attacks in the system which can be mitigated by the response r including the current attack.

 ISBN: 978-1-61208-398-8

Afterwards, we increase the response cost due to negative impact on the system expressed as in 5.

$$\sum_{s \in S_r} \omega(s), \tag{5}$$

where $\omega(s)$ denotes a function that calculates the importance of the security property s, which is affected by the response r and S_r is the set of security properties affected by the response r. The response impact is calculated in the same way as it is done for the attack, whereas the probability component is excluded as, unlike to intrusions, we can precisely define the impact of responses. Finally, we include the operational cost of the response r, denoted by function $opCost(r)$, in the same way as it is done for the attacks. Consequently, the response cost can be calculated as follows:

$$\begin{aligned} respCost(r) = &\sum_{a \in A} atCost(a) + \sum_{s \in S} \omega(s) + opCost(r) \\ &- respEff(r) \sum_{m \in M} atCost(m). \end{aligned} \tag{6}$$

Based on $respCost(r)$, we choose the response with the lowest cost value that is lower than the sum of the costs of existing and current attacks. If the response cost value is higher or equal than the overall cost, it means that the reaction can worsen the situation and additional investigation is required. The system resource state is considered healthy if for every security property of the system resource the following is true: there is no negative response impact and if there is an attack impact, it is mitigated by the deployed response(s).

V. Evaluation

We considered the following parameters as specified in [18] for the evaluation procedure: flexibility, dynamic, efficiency, ease of use, minimization of negative impact.

a) Flexibility: Flexibility of the proposed IRS is achieved by the system resources description method, which models dependencies between system resources as directed graph and is generally applicable for various environments. Response object properties can be changed as well to adapt priorities if it is required. The set of the system resources and response properties is not fixed and can be extended for additional flexibility.

b) Dynamic: Static IRS can be less efficient as they do not adapt to changes in the environment. Our proposed method tracks changes in the environment caused by attacks and responses and adapts to the current state of the system. This is achieved by getting feedback from the system after countermeasure deployment and consideration of already deployed responses.

c) Efficiency: The performance is one of the factors that affects the efficiency. We used the time for computing the results as measure of the performance. As our proposed IRS is intended to be used in small to medium sized enterprises (SMEs), the performance evaluation was performed with a desktop class computer with an Intel(R) Core(TM) i5-3330 CPU with 3 GHz and 8 GB RAM.

In Figures 1, 2 and 3, the average computation time for different numbers of attacks, responses and system resources is shown. During the analysis we varied the number of the analyzed type and fixed the number of the other types to 100. For example, if we perform an experiment by varying the number of system resources, the numbers of the attacks and responses was fixed to 100. The calculated computation time determines how long it took to process all generated attacks.

The computation time depends on the complexity of the system structure. The following system structure was used: a root resource is connected to all other resources, all dependent resources are connected to every node except the root node. Attacks are always performed on root node, so all resources are affected and are considered during computation. For each setup 10 experiments were performed and the average time was computed. For each experiment, the specified number of system resources and responses is created and then attacks are generated concurrently in multiple threads.

In Figure 1, it can be seen that the number of system resources affects computation time more than the size of the set of responses as depicted in Figure 2. Our proposed method aims on SMEs, which limits the number of system resources. So, if 1.6 seconds are required to process 100 attacks in the environment with 2500 system resources, on average it takes only 0.016 seconds to process one attack. We consider this time as acceptable as it significantly reduces the gap between intrusion detection and deployment of the response in comparison to classical manual approaches. Whenever, results for big environments are slower than ones demonstrated by Stakhanova, Strasburg, Basu, and Wong in [6], growth is near to linear.

For the results of the measurements as shown in Figure 3, the number of responses and attacked system resources was fixed to 100. The number of intrusions affects calculation time most of all, as was also concluded in [6]. The response selection procedure takes 47.34 seconds to process 2500 concurrent attacks, on average it takes only 0.019 seconds to process one attack. This result is close to the time required to process one attack in case of 2500 system resources in the environment. The rapid computational time escalation occurs as all attacks persisting in the system are involved in the countermeasure selection, because we also look for the responses which can minimize costs for mitigating not only the impact of the new attack, but also for mitigating all other attacks persisting in the system.

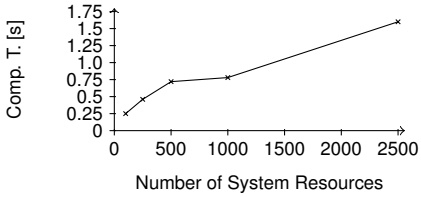


Figure 1. Performance (System Resources)

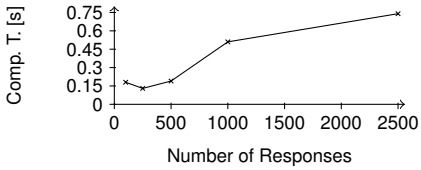


Figure 2. Performance (Responses)

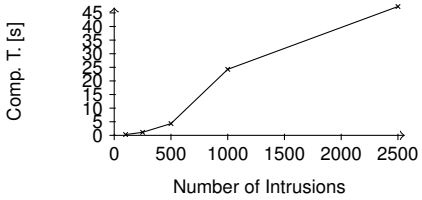


Figure 3. Performance (Intrusions)

d) Ease of use: One of the objectives for the research was to help system administrators of SMEs to deal with intrusions more efficiently. Our proposed system can be used both in automatic and semi-automatic modes. Consequently, system administrator can use an advantage to choose from the list of proposed responses and deployment locations. Additionally, we provide deployment and feedback mechanisms, which help in evaluating results of the deployment.

e) Minimization of negative impact: To illustrate minimization of the negative impact, we use an example assuming a simple environment with 2 hosts in the same sub-network. One of them contains web service for which availability is crucial, the second one contains FTP server for which integrity and confidentiality are important security properties (see Table I). Note that importance differs for the cases of partial and complete loss. Dependencies between resources with required conditions and possible effects are defined in the Table II. The set of the responses, including description of mitigation abilities and impact, is provided in the Table III. For simplification, we assume that the efficiency of the responses equals to 1 and that only one attack persists in the system. Additionally, the impact of a response is always either complete or none. We also assume that the attack impact and detection probabilities are equal to 1.0 and that the attack operational cost is equal to 0.5.

Consider the case, when sub-network instance is under attack and is intrusion entry point. The attack can affect confidentiality and integrity of the instance and the impact is complete. The probability of the impact on dependent resources is calculated as multiplication of the attack probability and probability of the transition to new state according to dependencies (see Table II). In accordance to (1), we compute the attack cost:

$$atCost(a) = 1 * (1 * 0.9 * 0.8 + 1 * 0.7 * 1.0 + 1 * 0.7 * 0.4 + 1 * 0.8 * 0.7 + 0.5) = 2.76.$$

The "isolate sub-network" is one of the possible responses. At first we set the value of the response cost to attack cost value: $respCost(r) = 2.76$. The impact of the response is equal to 2.1, as it affects the availability of the entire network, so we add this value to the cost. Additionally, we add the operational cost value. As a last step, we decrease the response cost due to efficiency against attack. Note that the efficiency coefficient equals to 1.0 for the sake of simplicity. According to (6), we get the response cost:

$$respCost(r) = 2.76 + 1.0 + 0.1 + 1.0 + 0.5 - 1 * 2.76 = 2.6.$$

Thereby, this response is already worth deploying. The "block port" reaction gives slightly better result $respCost(r) = 2.2$ due to lower operational cost. Additionally, we can deploy "isolate host" reaction to both hosts. Cost values will be $respCost(r)_{FTP} = 0.1 + 0.3 = 0.4$ for FTP server and $respCost(r)_{web} = 1.0 + 0.3 = 1.3$ for web server, with overall cost value $respCost(r) = 1.7$, which is better than both previous responses deployed on sub-network instance. Such cases are not considered by existing models. This example shows the importance of the deployment location. Nevertheless, the "block connection" response minimizes overall costs with response cost $respCost(r) = 0.1$.

Now, let us consider the case, when an attack partially affects confidentiality and integrity of the sub-network. The attack cost in this case is calculated as follows:

$$atCost(a) = 1 * (1 * 0.7 * 1.0 + 0.5) = 1.2.$$

The "network isolation" reaction has the following cost:

$$respCost(r) = 1.2 + 1.0 + 0.1 + 1.0 + 0.5 - 1 * 1.2 = 2.6.$$

The determined cost value is higher than the attack cost, thus it is better to keep the attack in the system as the deployment of this response will worsen the situation. The "block connection" response again gives the minimal cost $respCost(r) = 0.1$, while protecting all security properties. Another option is to disable an account. For this response, the cost will be $respCost(r) = 0.2$. In case of close cost values, the response efficiency coefficient, evaluated according to (3), allows making a decision based on the history of the response deployments in order to find the reaction with best chances to mitigate an attack.

TABLE I. SYSTEM RESOURCES

Sys. Res.	Confidentiality (Part./Compl.)	Integrity (Part./Compl.)	Availability (Part./Compl.)
Sub-network	0/0	0/0	0.7/1.0
FTP server	0.8/1.0	0.6/1.0	0/0.1
Web server	0/0.4	0.7/0.85	0.8/1.0

TABLE II. SYSTEM RESOURCES DEPENDENCIES

Sys. Res.	Dep. Sys. Res.	Dep. Cond.	Dep. Eff.	Prob.
Sub-net.	FTP	Conf. (Compl.)	Conf. (Part.)	0.9
		Int. (Part.)	Int. (Compl.)	0.7
		Avail. (Compl.)	Avail. (Compl.)	0.2
	Web	Conf. (Compl.)	Conf. (Compl.)	0.7
		Int. (Compl.)	Int. (Part.)	0.8
		Avail. (Part.)	Avail. (Compl.)	0.6

TABLE III. RESPONSES

Response	Confi. (Miti./Imp.)	Integr. (Miti./Imp.)	Avail. (Miti./Imp.)	Op. Cost
Isolate sub-net.	1/0	1/0	0/1	0.5
Isolate host	1/0	1/0	0/1	0.3
Block connection	1/0	1/0	1/0	0.1
Block port	1/0	1/0	0/1	0.1
Delay connection	0/0	0/0	1/0	0.2
Shutdown host	1/0	0/1	0/1	0.3
Disable account	1/0	1/0	0/0	0.2
Stop service	1/0	0/1	0/1	0.2

 ISBN: 978-1-61208-398-8

f) Limitations: While semi-automatic solutions are one step into the direction they still require human interaction and thus are slower than fully automated systems. Fully automated IRS, however, bring additional risks like erroneous responses or attackers misusing an IRS to trigger inadequate response deployment. Additional research on this topic is required.

The environment description process is a limitation of our proposed approach as it requires expert knowledge, is labor-intensive, and may be expensive. Our approach aims on SMEs, which typically lack resources and may not be able to afford personnel with expert knowledge in the field of information security. Notwithstanding, in case of not rapidly changing environment this limitation is not crucial, as the configuration is done once and reconfiguration is not required until any changes are introduced.

Assumptions according to the required information about the attack also imply limitations for the implementation in real environments. Our approach requires as precise and as detailed information as possible.

VI. Conclusion and Future Work

In this paper, we present an approach for optimizing the selection as well as the location of responses for mitigating attacks in computer networks. We have shown that multiple factors have to be taken into account. Existing solutions typically only consider subsets of factors that affect response selection mechanism in reality.

Our research focused on the practical application. Our proposed solution aims at balancing the costs of attacks versus the costs of countermeasures. We developed a model based on existing solutions that combines expert knowledge used for the resource assessment, increases accuracy of the selected response, and significantly reduces the gap between attack detection and response deployment with automated mechanism. Our approach allows to perform evaluation and adapt costs to the specific environments and security policies.

The concept was evaluated with a prototype implementation. The performance of the prototype showed acceptable results, even though the implementation can still be optimized. A possible optimization, e. g., is the pre-calculation of risk assessment values. Nevertheless, our prototype showed results that are sufficient for being used in real environments.

As future prospects, we plan to take more input parameters into account for the response selection. Additionally, we are going to research adaptable components in order to add flexibility and provide more optimized and efficient reactions to attacks and attack combinations.

Acknowledgment

This work was supported in part by the German Federal Ministry of Education and Research in scope of grant 16BY1201C. Responsible for the content are the authors.

References

[1] Symantec Corporation, “2013 trends,” Internet Security Threat Report 2014, vol. 19, April 2014.

[2] N. Stakhanova, S. Basu, and J. Wong, “A taxonomy of intrusion response systems,” International Journal of Information and Computer Security, vol. 1, no. 1/2, January 2007, pp. 169–184.

[3] N. B. Anuar1, M. Papadaki1, S. Furnell, and N. Clarke, “An investigation and survey of response options for intrusion response systems (irss),” Information Security for South Africa (ISSA), 2010, pp. 1–8.

[4] C. Mu and Y. Li, “An intrusion response decision-making model based on hierarchical task network planning,” Expert Systems with Applications: An International Journal, vol. 37, no. 3, March 2010, pp. 2465–2472.

[5] X. Zan, F. Gao, J. Han, X. Liu, and J. Zhou, “Nair: A novel automated intrusion response system based on decision making approach,” Information and Automation (ICIA), 2010, pp. 543–548.

[6] N. Stakhanova, C. Strasburg, S. Basu, and J. S. Wong, “Towards cost-sensitive assessment of intrusion response selection,” Journal of Computer Security, vol. 20, no. 2-3, 2012, pp. 169–198.

[7] Y. Wu and S. Liu, “A cost-sensitive method for distributed intrusion response,” Computer Supported Cooperative Work in Design (CSCWD), April 2008, pp. 760–764.

[8] N. Stakhanova, S. Basu, and J. Wong, “A cost-sensitive model for preemptive intrusion response systems,” Advanced Information Networking and Applications (AINA), May 2007, pp. 428–435.

[9] A. Ikuomola and A. S. Sodiya, “A credible cost-sensitive model for intrusion response selection.” in CASoN. IEEE, 2012, pp. 222–227.

[10] A. Shameli-Sendi, N. Ezzati-jivan, M. Jabbarifar, and M. Dagenais, “Intrusion response systems: Survey and taxonomy,” IJCSNS International Journal of Computer Science and Network Security, vol. 12, no. 1, January 2012, pp. 1–14.

[11] W. Lee, M. Miller, S. J. Stolfo, W. Fan, and E. Zadok, “Toward cost-sensitive modeling for intrusion detection and response,” Journal of Computer Security, vol. 10, 2002, pp. 5–22.

[12] A. Shameli-Sendi, J. Desfossez, M. Dagenais, and M. Jabbarifar, “A retroactive-burst framework for automated intrusion response system,” Journal of Computer Networks and Communications, vol. 2013, 2013.

[13] B. Foo, Y.-S. Wu, Y.-C. Mao, S. Bagchi, and E. Spafford, “Adepts: adaptive intrusion response using attack graphs in an e-commerce environment,” in Dependable Systems and Networks, 2005. DSN 2005. Proceedings. International Conference on, June 2005, pp. 508–517.

[14] Z. Wu, D. Xiao, H. Xu, X. Peng, and X. Zhuang, “Automated intrusion response decision based on the analytic hierarchy process,” Knowledge Acquisition and Modeling Workshop, December 2008, pp. 574–577.

[15] Standards for Security Categorization of Federal Information and Information Systems, “Fips pub 199 standards for security categorization of federal information and information systems,” Federal Information Processing Standarts Publication, 2004.

[16] I. Balepin, J. Rowe, and K. Levitt, “Using specification-based intrusion detection for automated response,” Recent Advances in Intrusion Detection, vol. 2820, 2003, pp. 136–154.

[17] P. Mell, K. Scarfone, and S. Romanovsky, “A complete guide to the common vulnerability scoring system version 2.0,” 2013.

[18] T. Toth and C. Kruegel, “Evaluating the impact of automated intrusion response mechanisms,” Computer Security Applications Conference, 2002, pp. 301–310.

 ISBN: 978-1-61208-398-8

Model for Cloud Computing Risk Analysis

Paulo F. Silva, Carlos B. Westphall, Carla M. Westphall
Networks and Management Laboratory
Post-Graduate Program in Computer Science
Federal University of Santa Catarina, Florianópolis, Brazil
e-mail:paulofernando@furb.br; westphal@inf.ufsc.br, carlamw@inf.ufsc.br

Mauro M. Mattos
Development and Transfer Technology Laboratory
Regional University of Blumenau, Blumenau, Brazil
e-mail: mattos@furb.br

Abstract – **Several risk analysis solutions have been proposed for cloud computing environments. But these solutions are usually centered on the Cloud Service Provider, have limited scope and do not consider the business requirements of the Cloud Consumer. These features reduce the reliability of the results of a cloud computing risk analysis. This paper proposes a model for cloud computing risk analysis in which responsibilities are not centered in the Cloud Service Provider. The proposed model makes the Cloud Consumer an active entity in risk analysis and includes the Information Security Laboratory entity. A prototype developed from the proposed model demonstrates performing a risk analysis in the cloud with shared responsibilities between the Cloud Service Provider, Cloud Consumer and Information Security Laboratory entities.**

Keywords – ISO 27005; cloud computing; risk analysis;

I. INTRODUCTION

Some of the challenges posed by cloud computing in the information security area are: identity management, virtualization management, governance and regulatory compliance, Service Level Agreement (SLA) and trust management, data privacy of the users and protection against external and internal threats [1]-[4].

Risk analysis [5] has been a strategy used to address the information security challenges posed by cloud computing, often addressing specific technical vulnerabilities or threats identification.

However, recent approaches on cloud risk analysis [6]-[12] did not aim at providing a particular architecture model for cloud environments, considering the entities involved and their responsibilities. Thus, the current models have the following deficiencies in their way of analyzing the risk of cloud computing environments:

- The deficiency in the adherence Cloud Consumer (CC) occurs when the entity responsible for defining impacts unaware of the technological environment and the CC business environment. In this case, the impact of this specification can disregard relevant scenarios for the CC or overestimate not relevant scenarios, thereby generating an incorrect risk assessment;
- The deficiency in the scope occurs when the selection of security requirements are performed by the Cloud Service Provider (CSP) itself or one without sufficient knowledge entity. The CSP can specify addict's security requirements in their own environment, thus defrauding the risk analysis results. Having an unprepared authority may specify requirements or insufficient disregard some important requirement, thus generating an incorrect risk analysis;
- The deficiency in the independence of results arises when the quantification of probabilities and impacts are performed by an entity that has an interest in minimizing the risk analysis results. For example, if the analysis is performed solely by the CSP. It can soften the assessment of requirements and impacts, thus generating a satisfactory result for the CC, but incorrect.

This paper proposes a model for performing risk analyzes in cloud environments that:

- Consider the participation of the CC entity in the performance of risk analysis, that is, allows an adherent risk analysis to CC's information security;
- Enabling the development of a risk analysis scope that is impartial to the interests of the CSP and to be developed by an entity with deep knowledge in information security;
- Does not have the centralized performance of risk analysis for the CSP entity, or to generate more independent results risks analysis possible in relation to the CSP interest, thus acting on the independence of disability results.

Therefore, the proposed model organizes the risk analysis in two phases: risk specification phase and risk evaluation phase. It also defines the entities involved in each phase and their responsibilities. Finally, the proposed model also provides a language for defining risk and a protocol for

 ISBN: 978-1-61208-398-8

communication between the entities involved in risk analysis.

The rest of this paper is as follows organized. Section 2 discusses related works on. The Proposed model is presented in Section 3. Section 4 describes the results and discussions. We conclude the paper and present future works in Section 5.

II. RELATED WORKS

Hale and Gamble [7] present a framework called SecAgreement that allows management of security metrics between CSPs and CCs. An SLA for cloud risk management is presented by Morin, Aubert and Gateau [8]. Ristov, Gusev and Kostoska [9] discuss risk analysis in cloud computing environments based on ISO 27001 and offers a model for security assessment in cloud computing. Chen, Wang and Wang [10] present an architecture that defines security levels from the risk of each CC service in the CSP.

Zech, Felderer and Breu [11] introduce a model for security testing in cloud computing environments based on risk analysis of these environments. Wang, Lin and Kuo [12] discuss risk analysis in cloud computing using intrusion techniques based on attack-defense trees and graphs.

Rot and Sobinska [13] discuss new information security threats specifically applied in cloud computing environments. Ristov and Gusev [14] present a security assessment of the main cloud environments open source, while Mirkovic [15] presents some security controls from ISO 27001 applied to cloud computing.

Ullah, Ahmed and Ylitalo [16] describe the Cloud Security Alliance (CSA) effort to inform security evaluation of automation in cloud services providers, the Cloud Audit, while Khosravani et al. [17] present a study of risk analysis in case of cloud computing, focusing on the importance of data security requirements that will be migrated to the cloud. Lenkala, Shetty and Siong [18] build upon the National Vulnerability Database (NVD) to identify vulnerabilities in cloud environments. Liu, Wu, Lu and Xiong [19] propose a model for information security risk analysis in virtual machines cloud computing environments, based on the ISO 27001, 27002 and 27005.

The related works presented above discuss the risk analysis on requirements or specific scenarios in cloud computing. The model proposed in this paper is different from the related works because it addresses an architecture for different risk scenarios in cloud computing, including discussion of the agents involved, communication protocol and language for description of the risks.

III. THE RACLOUD MODEL

This section presents the model for risk analysis proposed in the cloud, called RACloud – Risk Analysis for Clouds.

A. Risk Definition Language

The model provides a language for specifying risk, Risk Definition Language (RDL). The RDL is specified in XML and contains information about threats, vulnerabilities and information assets. This information is the basis for performing risk analysis in RACloud model.

The RDL allows specification of three different types of records: threats, vulnerabilities and information assets. Figure 1 shows an example of specifying vulnerability records, which are two specified vulnerabilities from the Common Vulnerabilities and Exposures (CVE).

Each record contains information RDL header with Id, source and version of the XML file and registry information risk (threats, vulnerabilities or information asset) with Id, description, category and Web Service Risk Analysis (WSRA).

The WSRA is a web service responsible for evaluating the record of risk (threat, vulnerability and asset information). It is also responsible by quantifying the risk as shown in Section III-C.

```
<RDL type="ISL" id="1299">
    <source>LRG-UFSC</source>
    <version>4.5.1a</version>
    <description>...</description>
    <vulnerabilities>
        <item id="129">
            <description>Cipher protocol weak</description>
            <category>service</category>
            <wsra>http://lrg.ufsc.br:8095/evaluate129</wsra>
        </item>
        <item id="239">
            <description>Clear text password</description>
            <category>service</category>
            <wsra>http://lrg.ufsc.br:8095/evaluate239</wsra>
        </item>
    </vulnerabilities>
</RDL>
```

Figure 1. An RDL especification of vulnerabilities.

The RDL records are used by the components of the model RACloud (Section III-B) during phases of risk specification (Section III-D) and risk assessment (Section III-E).

B. Architectural Components

The RACloud model shares the responsibility of risk analysis between four distinct entities: RAH - Risk Analysis Host, ISL - Information Security Laboratory, CSP - Cloud Service Provider and CC - Cloud Consumer. These entities relate to different components at different times.

The RAH entity has responsibility for the host connection and core layers, formed by components Conn ISL, Conn CC, Conn CSP, Agent Manager, RDL Manager and Analysis Manager (Figure 2).

The components Conn ISL, Conn CC and Conn CSP are interfaces for communication with other components distributed respectively between the entities ISL, CC, CSP.

The Agent Manager component is responsible for managing the registration of CC, CSP and ISL entities in RACloud model. The RDL Manager component is responsible for managing and storing the records of defining risks. And Analysis Manager component is responsible for performing the risk assessment.

 ISBN: 978-1-61208-398-8

The ISL is a laboratory entity or group specializing in information security, its responsibility is to specify the RDLs vulnerabilities and threats, in addition to their WSRAs. This entity hosts the ISL Agent and WSRA Evaluator components. The component ISL Agent is responsible for registering the ISL in RACore and publish its RDLs. The WSRA Evaluator component is responsible for performing assessments of threats and vulnerabilities described in RDLs.

The CSP represents the entity's own cloud service provider aim of risk analysis. This entity hosting the CSP Agent and WSRA Proxy components. The CSP Agent component is responsible for registering the CSP in RACore and subscribe to RDLs, which the CSP aims to be analyzed. The WSRA Proxy component is responsible for collecting information from the CSP and make the call of WSRAs.

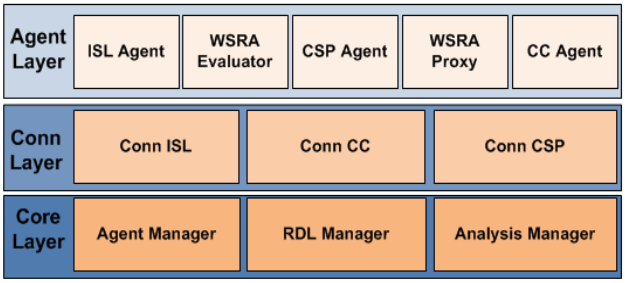


Figure 2. Model Layers RACloud.

The CC entity is the CSP's customer, hosting their information assets in the cloud and want to know which one is exposed to risk in relation to its CSP. This entity hosts the CC Agent component. This component is responsible for registering the CC in RACloud and initiate risk analysis.

C. *Risk Modeling*

Information assets, threats and vulnerabilities are the basic elements of a risk analysis of information security. RACloud in these model elements are defined by CC and ISL entities. Variables modeling of risk posed information assets, threats and vulnerabilities are shown in Table I.

TABLE I. BASIC ELEMENTS

Symbol	Description
T_x	Treat defined by ISL "x"
A_y	Information Asset defined by CC "y"
V_z	Vulnerability defined by ISL "z"

In the risk analysis, functions are applied to the information assets, threats and vulnerabilities, with the aim of analyzing their impact, exposure and disability, respectively. The functions for allocating degree of impact, degree of exposure and degree of disability are represented according to Table II.

TABLE II. FUNTIONS OF ANALYSIS

Symbol	Description
$eaf(T_x,w)$	Exposure analysis function of T_x on CSP "w"
$iaf(A_y)$	Impact analysis function of A_y
$daf(V_z,w)$	Deficiency analysis function of V_z on CSP "w"

The analysis functions represented in Table II result in the calculation of the degree of impact, degree of exposure and degree of disability. The three variables are represented in RACloud as described in Table III.

TABLE III. VIABLES OF ANALYSIS

Symbol	Description
$DE_{T,x,w}$	Degree of Exposure related with T_x and w. $eaf(T_x,w)=DE_{T,x,w}$
$DI_{A,y}$	Degree of Impact related with A_y. $iaf(A_y)=DI_{A,y}$
$DD_{V,z,w}$	Degree of Deficiency related with V_z and w. $daf(V_z,w)=DD_{V,z,w}$

A risk event is the relationship of a threat with a vulnerability. This relationship is established in RACloud through a correlation function of the event. From the risk events are calculated the probabilities of occurrence of the event, based on the degree of exposure and the degree of disability. The modeling related events and probabilities is presented by Table IV.

TABLE IV. PROBABILITY CALCULATION

Symbol	Description
$E_{T,V}$	Event relating T with V
$\alpha(T_x,V_z)$	Function correlating T and V $\alpha(T_x,V_z)=E_{T,V}$
$fp(E_{T,V})$	Function of probability of $E_{T,V}$ $fp(E)=(DE_{T,x,w}+DD_{V,z,w})/2$, or, $fp(E)=matrix(DE_{T,x,w},DD_{V,z,w})$
P_E	Probability of $E_{T,V}$ $fp(E_{T,V})=P_E$

From the probability of risk events and the degree of impact on information assets, it is possible to calculate the risk of a particular event on a particular information asset. The relationship between risk events and information assets are given by a function correlation risk. The modeling related to the correlation of risk and the final calculation of risk is presented by Table V.

TABLE V. RISK CALCULATION

Symbol	Description
$R_{E,A}$	Risk relating E and A
$\beta(E,A_y)$	Function correlating E and A_y $\beta(E,A_y)=R_{E,A}$
$raf(R_{E,A})$	Risk analysis function of $R_{E,A}$ $raf(R_{E,A})=(P_E+ DI_{A,y})/2$ or $raf(R_{E,A})=matrix(P_E,DI_{A,y})$
$DR_{E,A}$	Degree of risk related with $R_{E,A}$ $raf(R_{E,A})=GR_{E,A}$

D. *Specification Phase*

In the risk specification phase, RACloud model of the threats (T_x), vulnerabilities (V_z) and information assets (A_y) part of risk analysis is defined.

Figure 3 illustrates the flow of interactions between components of the model RACloud specification phase risk. Initially each agent must register with the Agent Manager component (Figure 3 -a, b, c). After it registed ISL has the responsibility to specify threats and vulnerabilities of cloud computing environments and develop RDLs and WSRAs to these threats and vulnerabilities. The vulnerabilities and threats WSRAs to match functions eaf(T_x,w) e daf(V_z,w) of the risk modeling, respectively.

After developing their RDLs and WSRAs ISL exports the records of RDLs for the RDL Manager (Figure 3 -d) component and publishes WSRAs (Figure 3 -e) so they can be called by the CSP in the evaluation phase.

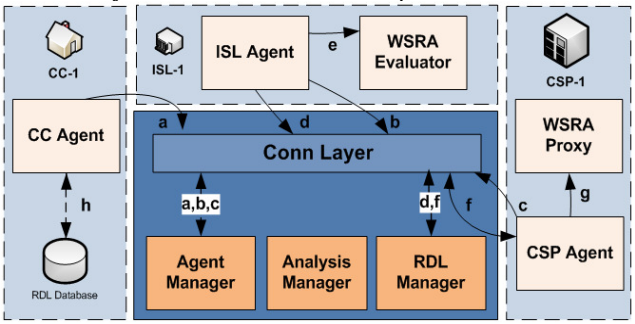


Figure 3. Specification time.

The performance of the CSP specification phase of risk is to import the RDLs recorded by ISL (Figure 3 -f.) and implement the Proxy WSRAs to call WSRAs of the evaluation phase (Figure 3 -g).

The identification of threats and vulnerabilities, is the responsibility of the ISL and the call of WSRAs, is the responsibility of the CSP, but the definition of information assets and quantification of impact on these assets is the responsibility of the CC. Because CC entity is most adequate for the express the potential loss in the event of an incident. Thus, the responsibility of CC Agent on phase specification risk is to build a database of information assets RDLs (Figure 3 -h).

E. Evaluation Phase

In the risk evaluation phase, it ocorrs the call of the functions eaf(T_x,w), daf(V_z,w) e iaf(A_y), and quantifying the variables $E_{T,V}$, P_E and $R_{E,A}$ defined in risk modeling.

The evaluation begins with the CC Agent informing the CSP to be analyzed (Figure 4 -a). From this component Analysis Manager obtains information from the CSP (Figure 4 -b) and queries the registered RDLs (Figure 4 -c).

Based on information obtained from CSP and RDL, Analysis Manager component starts and will evaluation threats and vulnerabilities. To do so, makes the invocation of CSP Agent. Then there is the collection of information about threats and vulnerabilities through WSRAs Proxy and the assessment of that information through WSRAs ISL. Then WSRAs ISL make quantification of the variables $DE_{T,x,w}$ and $DD_{V,z,w}$ (Table II) and return these values to Analysis Manager component (Figure 4 -d).

After quantifying all the threats and vulnerabilities associated with RDLs defined in CSP, begins to quantify the impacts defined by the CC. Therefore, the Analysis Manager component invokes the CC Agent for the degree of impact of their information assets (Figure 4 -e). With the return of CC Agent Analysis Manager component defines the value of the variables $DI_{A,y}$.

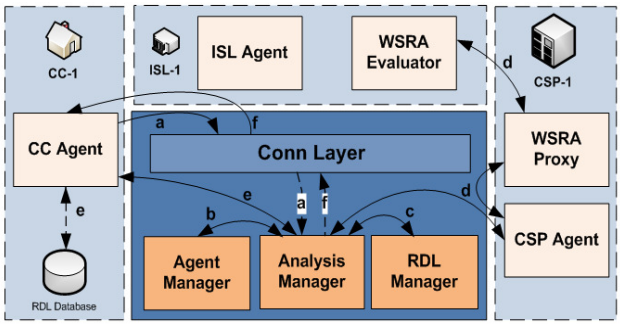


Figure 4. Evaluation time.

Once obtained the values of $DI_{A,y}$, $DE_{T,x,w}$ e $DD_{V,z,w}$ for all information assets, threats and vulnerabilities defined in RDLs, Analysis Manager component starts the calculation of the variables $E_{T,V}$ and P_E and through the functions $\alpha(T_x,V_z)$ and $fp(E_{T,V})$. This process results in a list of possible events, or which may threats and vulnerabilities which exploits, and the respective probability of each event.

Finally, the Analysis Manager component does the calculation of the variables $R_{E,A}$ and $DR_{E,A}$, through the functions $\beta(E,A_y)$ and $raf(R_{E,A})$ respectively. The result of this process is the ratio of risk items, ie valid relation between events and information.

After the calculation of all risk items ($R_{E,A}$) and their degrees of risk ($DR_{E,A}$) the result is returned to the CC Agent for it to take decisions on the acceptance or not of risk found in their CSP (Figure 4 -f).

IV. RESULTS AND DISCUSSION

For testing purposes and discussion, we developed a prototype RACloud model as presented in Section III. From the prototype were performed phases of risk specification and risk evaluation in a controlled environment for testing.

In the risk of specification phase (Section III-D), were specified 20 RDL records vulnerabilities and 20 RDL records threats and 10 RDL records of information assets. The RDL records of threats and vulnerabilities were specified as threats and vulnerabilities found in CVE -. Common Vulnerabilities, Exposures. Also WSRAs and WSRAs Proxy have been developed for the 40 records of threats and vulnerabilities specified.

In the risk evaluation phase (Section III-E), the WSRAs Proxy and WSRAs were performed, generating the DD and DE values for each vulnerability and threat record,

 ISBN: 978-1-61208-398-8

respectively. The records of vulnerabilities and threats were correlated by Analysis Manager component generates 20 events, which were correlated with the records of information assets, generating 20 risk scenarios.

Figure 5 shows the result of calculation of variables DE, DD, P, DI and DR for the 20 risk scenarios (R1 to R20) specified in the prototype.

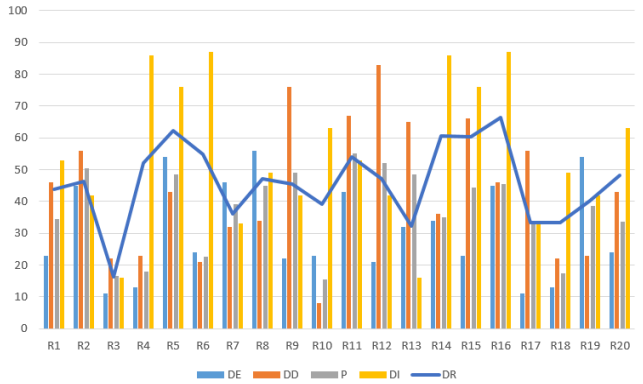


Figure 5. Evaluation of risk.

The lower risk identified was the R3 risk scenario, with risk of 16.25%. This scenario specifies as information asset the file transfer service, as vulnerability the unencrypted password and as threat the unauthorized access.

The greatest risk identified was the risk scenario R16, with risk of 66.25%. This risk scenario specifies as information asset the e-mail service, as vulnerability the weak encryption protocol and as threat the DDoS.

Figure 6 presents the results of the risk assessment generated by RACloud model prototype for the risk scenarios R3 and R16. For each risk scenario is possible to observe the results of probability and risk variables. You can also see a brief description of the items threats, vulnerabilities and information assets and the value of their respective variable degree of exposure, degree of deficiency and degree of impact.

With the risk analysis of the resulting information the CC may decide to allocate or not their information assets in a given CSP, or remove their systems of a CSP to present great risks.

The proposed model aims to reduce the three major deficiencies presented by current models of cloud risk analysis: deficiency in scope, deficiency in the adherence and deficiency in independence of results.

The reduction deficiency in the adherence occurs when the proposed model includes the CC as a key entity in the risk analysis process. In the model RACloud, the CC entity acts in active mode on risk analysis, defining information assets and quantifying impacts on these assets.

The CC is the entity most apt to define the impacts, it is the entity that best knows the relevance of each information asset within its area of operation. Therefore, it is CC's responsibility to say what the impact will be whether a system file or database has its integrity, confidentiality or availability impaired. The CSP and ISL entities have no automy to identify or quantify impacts on information assets, because they are not experts in CC business area.

```
<RDL Id="248" type="RISK">
   <source>RACloud-LRG</source>
   <version>5a</version>
   <description>...</description>
   <cc_id>consumerCC</cc_id>
   <csp_id>testCSP</csp_id>
   <risks>
      <item id="3">
         <probability>16.25</probability>
         <risk>42</risk>
         <informationasset DI="16">File transfer service</informationasset>
         <vulnerability DD="22">Clear text password</vulnerability>
         <treat DE="11">Unauthorized Access</treat>
      </item>
      <item id="16">
         <probability>45.5</probability>
         <risk>66.25</risk>
         <informationasset DI="87">Email service</informationasset>
         <vulnerability DD="46">Cipher protocol weak</vulnerability>
         <treat DE="45">DDos</treat>
      </item>
   </risks>
</RDL>
```

Figure 6. Result of risk.

The RACloud model works to reduce the deficiency in scope in that it introduces the ISL entity. As the ISL an entity specialized to information security is the entity best placed to define security requirements, threats and vulnerabilities (specification of RDLs) and set as the threats and vulnerabilities should be quantified (specification of WSRAs).

The reduction of deficiency in the independence of the results is the fact that the model RACloud the CSP has more restricted responsibilities than in the models traditionally presented by related work.

Traditionally, the CSP is responsible for defining security requirements and the tests that are applied to risk assessment of their own environment. In this scenario the risk assessment may be biased to the CSP. Including the ISL entity removes responsibilities traditionally assigned to the CSP, as identification and quantification of threats and vulnerabilities, thus making it more reliable the result of risk analysis.

The proposed model allows multiple ISLs act in the definition of RDLs and WSRAs together. Thus the risk definitions can come from different sources and can be constantly updated dynamic and collaborative way, forming a risk settings based on extensive and independent cloud.

The way WSRAs are specified is also a feature that impacts the improvement scope. The use of Web Services to specify security requirements allows them to be platform independent and can be ordered by any CSP. It also allows the use of a wide variety of techniques for quantification of threats and vulnerabilities, because the limit is defined only by the programming language chosen for implementation of WSRA.

The related works of cloud risk analysis did not consider the role of CC entity in the risk analysis. These works usually aim on the vulnerability assessment by the CSP itself, without considering the impact that the vulnerability will cause on the different CC information assets. By assigning the responsibility for identifying and quantifying

 ISBN: 978-1-61208-398-8

the impact of the CC are sharing the risk variables among different entities, so the responsibility for the quantification of risk analysis variables is not centralized in one specific entity.

The CSP is the entity that will be the analyzed then it doesn't have the autonomy to set any of the values of risk analysis, as this could make unreliable risk analysis. The role of CSP is only inform the data requested by ISL, so that ISL itself makes the quantification of security requirements.

With RACloud model CC can perform analyzes in several CSPs before deciding to purchase a cloud computing service. The CC can also carry out regular reviews of your current provider and compare them with other providers, opting for changing its CSP.

V. CONCLUSION

This paper presented a model for risk analysis in cloud computing environments.

The proposed model changes the generally current paradigm in research on cloud risk analysis, in which the CSP entity is responsible for the specification of security requirements and analysis of these requirements in its own environment, so the only entity responsible for the results risk analysis.

To reduce excess CSP responsibility for risk analysis, the proposed model includes two new entities with active participation in risk analysis, the CC entity and the ISL entity.

The model presented in this paper is an initiative of the CC itself can perform risk analysis on its current or future CSP. And that this risk analysis is adherent, comprehensive and independent of the CSP interests.

The characteristics presented in this paper are intended to generate a more reliable risk analysis for CC, so that it can choose its CSP based on more consistent information, specified and analyzed by an exempt entity interests, ISL.

Several papers on cloud computing indicate lack confidence CC in relation to the CSP as a great motivator for not acquiring cloud computing services. An independent risk analysis can act to reduce this mistrust and promote the acquisition of cloud computing services.

The prototype and the results show the specification and implementation of an adherent risk analysis, comprehensive and independent, because the analysis is not centered in the CSP. The identification and quantification of threats and vulnerabilities can be performed by many security laboratories and the impact on the information assets is defined by the CC itself.

Several future works can be developed from the RACloud model. There is a need to extend this work to suggest the controls or countermeasures for CSPs can mitigate its risks. Searches can be developed on the reliability of the data reported by the CSP to the ISL for risk analysis and the specification of risk definition language can be further explored in specific researches.

REFERENCES

[1] M. K. Srinivasan, K. Sarukesi, P. Rodrigues, M. S. Manoj, and P. Revathy, "State-of-the-art cloud computing security taxonomies: a classification of security challenges in the present cloud computing environment". ICACCI '12: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, August 2012, pp. 470-476.

[2] H. Yu, N. Powell, D. Stembridge and, X. Yuan, "Cloud computing and security challenges". ACM-SE '12: Proceedings of the 50th Annual Southeast Regional Conference, March 2012, pp. 298-302.

[3] K. Ren, C. Wang and Q. Wang, "Security Challenges for the Public Cloud," Internet Computing, IEEE, vol.16, no.1, Jan.-Feb. 2012, pp. 69-73, doi: 10.1109/MIC.2012.14, retrieved: March, 2015.

[4] B. Grobauer, T. Walloschek and E. Stocker, "Understanding Cloud Computing Vulnerabilities," Security & Privacy, IEEE , vol.9, no.2, March-April 2011, pp. 50-57, doi: 10.1109/ MSP.2010.115.

[5] ISO/IEC 27005:2011, Information Security Risk Management. [Online]. Available: http://www.iso.org, retrieved: March, 2015.

[6] J. Zhang, D. Sun and D. Zhai, "A research on the indicator system of Cloud Computing Security Risk Assessment," Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), 2012 International Conference on , vol., no., June 2012, pp.121,123, 15-18 doi: 10.1109/ ICQR2MSE. 2012.6246200.

[7] M. L. Hale, and R. Gamble, "SecAgreement: Advancing Security Risk Calculations in Cloud Services," Services (SERVICES), 2012 IEEE Eighth World Congress on , vol., no., June 2012, pp.133-140, 24-29, doi: 10.1109/SERVICES.2012.31.

[8] J. Morin, J. Aubert, and B. Gateau, "Towards Cloud Computing SLA Risk Management: Issues and Challenges," System Science (HICSS), 2012 45th Hawaii International Conference on , vol., no., pp.5509-5514, 4-7 Jan. 2012 doi: 10.1109/HICSS.2012.602.

[9] S. Ristov, M. Gusev, and M. Kostoska, "A new methodology for security evaluation in cloud computing," MIPRO, 2012 Proceedings of the 35th International Convention , vol., no., May 2012, pp.1484-1489, 21-25.

[10] J. Chen, Y, Wang, and X. Wang, "On-Demand Security Architecture for Cloud Computing," Computer, IEEE, vol.45, no.7, July 2012, pp.73,78, doi: 10.1109/MC.2012.120.

[11] P. Zech, M. Felderer, and R. Breu, "Towards a Model Based Security Testing Approach of Cloud Computing Environments," Software Security and Reliability Companion (SERE-C), 2012 IEEE Sixth International Conference on , vol., no., June 2012, pp.47,56, 20-22 doi: 10.1109/SERE-C.2012.11.

[12] P. Wang, W. Lin, P. Kuo, H. Lin and, T. Wang, "Threat risk analysis for cloud security based on Attack-Defense Trees," Computing Technology and Information Management (ICCM), 2012 8th International Conference on, vol.1, no., April 2012, pp.106-111, 24-26.

 ISBN: 978-1-61208-398-8

[13] A. Rot, and M. Sobinska, “IT security threats in cloud computing sourcing model”, Computer Science and Information Systems (FedCSIS), 2013, Federated Conference on, Publication Year: 2013, pp. 1153- 1156.

[14] S. Ristov, and M. Gusev. “Security evaluation of open source clouds”, EUROCON, 2013 IEEE, Digital Object Identifier: 10.1109/EUROCON. 2013.6624968, Publication Year: 2013, Page(s): 73- 80.

[15] O. Mirkovic, “Security evaluation in cloud”, Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on, Publication Year: 2013 , Page(s): 1088-1093.

[16] K. Ullah, A. Ahmed, and J. Ylitalo. “Towards Building an Automated Security Compliance Tool for the Cloud”. Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on. Digital Object Identifier: 0.1109/TrustCom. 2013.195. Publication Year: 2013, Page(s): 1587- 1593.

[17] A. Khosravani, Nicholson, B., and Wood-Harper, T., “A case study analysis of risk, trust and control in cloud computing”, Science and Information Conference (SAI), 2013, Publication Year: 2013, Page(s): 879- 887.

[18] S. R. Lenkala, Shetty, S., and Kaiqi Xiong. “Security Risk Assessment of Cloud Carrier”. Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on, Digital Object Identifier: 10.1109/CCGrid.2013.28, Publication Year: 2013, Page(s): 442- 449.

[19] S. Liu, J. Wu, Z. Lu, and H. Xiong, “VMRaS: A Novel Virtual Machine Risk Assessment Scheme in the Cloud Environment”, Services Computing (SCC), 2013 IEEE International Conference on, Digital Object Identifier: 10.1109/SCC.2013.12, Publication Year: 2013, Page(s): 384-391.

 ISBN: 978-1-61208-398-8

Classifying Anomalous Mobile Applications Based on Data Flows

Chia-Mei Chen
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan
Email: cchen@mail.nsysu.edu.tw

Yu-Hsuan Tsai
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan
Email: t0336470@gmail.com

Gu-Hsin Lai
Department of Information Management
Chinese Culture University
Taipei, Taiwan
Email: lgx4@ulive.pccu.edu.tw

Sheng-Tzong Cheng
Department of Computer Science and Info. Engineering
National Cheng Kung University
Tainan, Taiwan
Email: stcheng@mail.ncku.edu.tw

***Abstract*—Mobile security becomes more important as users increasingly rely on the portable network devices. The security consultant firms indicate that the amount of mobile malware increases every year at a fast speed. Therefore, fast detecting mobile malware becomes an important issue. By applying reverse engineering techniques, a source code extraction module produces data flow information from the mobile application executable. The proposed static analysis-based detection system analyzes the data flow of the target software and it identifies if a data flow might leak sensitive data. The experimental results show that the proposed detection system can identify mobile malware efficiently.**

Keywords- mobile security, malware detection, static analysis.

I. INTRODUCTION

Mobile users get used to downloading various mobile applications on the mobile devices for business as well as leisure purposes. Therefore, confidential information is stored in the mobile devices which become the new target for financial gain. Juniper Networks study [1] states that 92% of mobile malware targets the Android platform, as it has the highest market share. Tread Micro [3] reports that seventeen pieces of malware had already been downloaded seven hundred thousand times before they were removed and half of mobile malware involve unauthorized text message sending or network access. F-Security report [5] concludes that mobile malware are mostly profit oriented and security might be the primary concern for mobile users. The number of apps increases dramatically in the markets and an efficient mobile malware detection is demanded.

Commercial mobile malware detection solutions such as BullGuard Mobile Security and Lookout Mobile Security adopt signature-based approach [2] and the detection rate relies on the malware signature repository. For fast growing mobile malware, hackers have a chance to compromise mobile users before the signature is developed [14]. Hence, an alternative solution should be developed to detect unknown mobile malware.

In this research, the proposed detection system develops a feature selection method combining genetic algorithm and data flow analysis, where genetic algorithm reduces the number of features and data flow analysis shows the relationship between API calls and system commands. This research conducted a preliminary study analyzing collected mobile apps and malware and discovered that apps authors might obfuscate the codes by replacing variable names into meaningless strings but the API calls and system commands would not be altered. To steal privacy information, certain API calls and system commands would be invoked. Therefore, the proposed detection method considers the API calls and system commands as key attributes. Based on our preliminary study, the possible sequences of the API calls and system commands are huge. Therefore, genetic algorithm is applied to build efficient threat patterns of the API call and system command invocation. The proposed detection method can identify unknown malware which matches the malicious behaviors found.

The structure of the paper is organized as follows. The literature review is studied in Section II. Section III describes the proposed classification method, followed by performance evaluation in Section IV. The conclusion remarks are drawn in Section V.

II. RELATED WORK

Dynamic analysis and static analysis [21] are common approaches used for malware detection. Dynamic analysis consumes more resources and computation time, while static analysis requires source code or reverse engineering.

Bhaskar Pratim et al. [9] proposed an approach which analyzes the risk of an app based on permission. The approach is limited to the official Google Play market, but most malware resides in the third party markets. Francesco Di Cerbo et al. [11] applied Apriori algorithm to identify common subsets of permissions used by the benign apps.

As app writers may produce over-privileged mobile software [27][28], permission based approach might not be enough to identify mobile malware. Some malware even

 ISBN: 978-1-61208-398-8

conducts malicious behaviors without permission [29]. Permission based mobile malware detection has drawbacks [20] and is not efficient.

William et al. [30] built an Android sandbox by modifying Android's source code. The sandbox traces the data flows of the sensitive data, such as IMEI or DeviceId, which appears in text messages or network connection. This method is designed for security researchers monitoring data flows in the mobile devices but not suitable for detecting mobile malware.

Shabtai et al. [14] proposed a detection system applying knowledge-based and temporal abstraction method to detect unknown malware. Temporal patterns of mobile devices are established from history events such as app installation and the number of text message sent out. A monitored event without user interaction is regarded as abuse. In the practical cases, users tend to press OK when using an app and hackers could apply social engineering tactics to circumvent such restriction.

Wu et al. [10] proposed a malware classification method which combines several types of features: permission and component information from Manifest file, information of intent, API calls and communication between components from source code. K-mean algorithm and expectation–maximization algorithm are applied to classify the mobile applications. Yerima et al. [7] proved that API calls and system calls are efficient for distinguishing malware and benign applications and Bayesian classifier is adopted to classify malwares and benign applications.

The above mentioned classification approaches do not provide the cause of malicious behaviors and might confuse users. The literature indicates that API calls and system calls are efficient and the invocation ordering is useful for defining malicious behaviors. Therefore, the proposed detection system develops an efficient feature selection method to identify efficient features and build the invocation sequences used by malware. With reduced feature sets, the proposed detection reduces the detection time without detection performance loss.

III. PROPOSED SYSTEM

The literature review and our preliminary study indicate that obfuscated software replacing variable names to meaningless strings makes static analysis based detection hard and each piece of software has unique invocation sequence. API call and system command invocation represents the behaviors of a piece of software. The distinct sequences of the invocations could increase large as the number of malware raises. Therefore, the proposed detection system develops a feature selection method which applies genetic algorithm to reduce the number of feature sets and build efficient threat patterns of API call and system command invocation.

The proposed system consists of three processes, reverse engineering, threat pattern building, and detecting processes as shown in Figure 1.

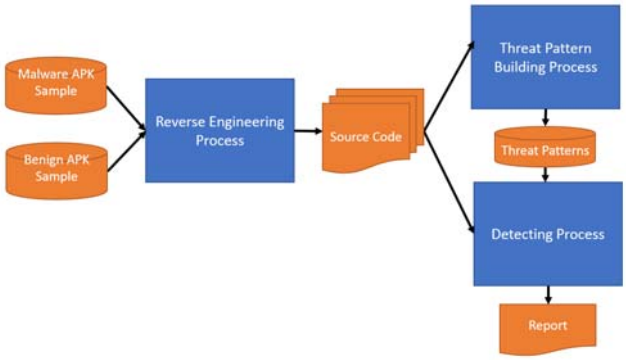


Figure 1. System architecture

Three tools, APKTool, dex2jar, and JAD, are applied for reversing app's APK files into source code. APKTool produces the .dex files from the apk files; dex2jar transforms the .dex files into a set of the .class files; and decompiler JAD converts the .class files into the .jad files which are the Java source code of the APK files. Source code provides valuable information. API calls and system commands can be retrieved.

Threat patterns are sequences of API calls and system command invocations. Some API calls and system commands are invoked by both malicious and benign apps, and are not distinguishable features for malware detection. Therefore, in the threat pattern building process, the feature selection module eliminates common calls and commands used by two types of the mobile applications.

Feature Set Reduction by Genetic Algorithm

Many API calls and system commands were found in the test dataset; therefore, the number of the possible combinations of the invocations is huge. The feature sets grow up as the number of invocation sequences increases. In this study, genetic algorithm is applied to select a suboptimal set of invocations which can distinguish mobile malware from the normal apps. The goal of the proposed classification system is to maximize the detection rate which is measured by true positive rate and precision in this study. Hence, the proposed fitness function is defined by the detection performance measurements mentioned above: true positive rate + precision.

IV. SYSTEM EVALUATION

The mobile apps for evaluation were extracted from Android Malware Genome Project [26] and Google Play Market. This study assumes that the chance of a malicious and popular app which can survive in Google Play market for over three month is low.

 ISBN: 978-1-61208-398-8

Table I. DETECTION RESULTS.

Family	No of apps in the family	No. of detected malware	True positive
ADRD	22	22	100.00%
AnserverBot	186	187	99.47%
Asroot	7	8	87.50%
BaseBridge	115	122	94.26%
BeanBot	8	8	100.00%
Bgserv	9	9	100.00%
CoinPirate	1	1	100.00%
CruseWin	2	2	100.00%
DogWars	0	1	0.00%
DroidCoupon	0	1	0.00%
DroidDeluxe	1	1	100.00%
DroidDream	15	16	93.75%
DroidDreamLight	46	46	100.00%
DroidKungFu1	33	34	97.06%
DroidKungFu2	30	30	100.00%
DroidKungFu3	309	309	100.00%
DroidKungFu4	96	96	100.00%
DroidKungFuSapp	3	3	100.00%
DroidKungFuUpdate	1	1	100.00%
Endofday	1	1	100.00%
FakeNetflix	0	1	0.00%
FakePlayer	0	6	0.00%
GGTracker	1	1	100.00%
GPSSMSSpy	0	6	0.00%
GamblerSMS	1	1	100.00%
Geinimi	69	69	100.00%
GingerMaster	4	4	100.00%
GoldDream	47	47	100.00%
Gone60	0	9	0.00%
HippoSMS	2	4	50.00%
Jifake	0	1	0.00%
KMin	52	52	100.00%
LoveTrap	1	1	100.00%
NickyBot	1	1	100.00%
NickySpy	0	2	0.00%
Pjapps	57	57	100.00%
Plankton	11	11	100.00%
RogueLemon	2	2	100.00%
RogueSPPush	9	9	100.00%
SMSReplicator	1	1	100.00%
SndApps	10	10	100.00%
Spitmo	1	1	100.00%
Tapsnake	0	2	0.00%
Walkinwat	0	1	0.00%
YZHC	22	22	100.00%
Zitmo	0	1	0.00%
Zsone	12	12	100.00%
jSMSHider	16	16	100.00%
zHash	11	11	100.00%
Total	1215	1259	96.51%

The detection results are shown in Table I; the proposed system has the detection rate of 96.5%. The proposed detection method might have false negative on small size malware families, as the threat patterns used by them improve insignificantly on fitness function of the genetic algorithm. As for false positive, 119 benign samples out of 1,259 were classified as malicious. Some misclassified samples have root threat. For example, com.estrongs.android.pop.cupcak is one of the applications that being detected has root threat. As shown in Figure 2, the description of this application indicates that it requires root access. Most misclassified benign apps were detected as malwares because of data thief threat. For example, data synchronization app, com.gozap.labi.android copies information stored in the mobile device and sends to somewhere. Other misclassifications were caused by adware which sends out device ID for advertisement purpose [19]. Therefore, the results conclude that the proposed detection system can detect malware efficiently.

Figure 2. Misclassified benign app (snapshot of com.estrongs.android.pop.cupcak in Google Play) requires root access.

V. CONCLUSIONS

Mobile devices are widely used in our daily work and leisure time. The security surveys and reports demonstrate that hackers have shifted the attack target to mobile users and mobile malware increases each year. Signature based detection is not suitable for fast growing and changing mobile malware.

Static analysis is suitable for analyzing fast growing mobile malware. This study proposes a static analysis based detection method which identifies efficient feature sets from the API calls and system commands. Two phases of feature set reductions are developed and the experimental results show that the proposed detection using the feature selection method performs efficiently with the detection rate of 96.5%.

Further evaluation and investigation should be made to compare the proposed static analysis approach with signature based detection method and to analyze the process time required by the proposed system in the reverse engineering and model training phases.

Static analysis might have limitations. Malware with botnet capability which receives and executes attack commands from command and control server might not be detectable from static analysis.

The reverse engineering tools and techniques used in this study can be improved to extract better quality of source code. Some applications use NDK (Native Development Kit) which allows to develop functions in language C and to extend invocation via JNI (Java Native Interface). The C functions are compiled into share object (.so file) and hard to decompile back to the source code. The software which invokes malicious functions in C requires better detection

 ISBN: 978-1-61208-398-8

and reverse engineering methods to identify the anomalous behaviors.

REFERENCES

[1] Juniper networks, "Juniper networks Mobile threat Center Third Annual Mobile threats report," retrieved March 1, 2015 from http://www.juniper.net/us/en/local/pdf/additional-resources/3rd-jnpr-mobile-threats-report-exec-summary.pdf.

[2] Shuaifu Dai, "Behavior-Based Malware Detection on Mobile Phone" The 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), 2010.

[3] TrendMicro, "Android Malware: How Worried Should You Be?" retrieved on March 1, 2015 from http://blog.trendmicro.com/trendlabs-security-intelligence/android-malware-how-worried-should-you-be/.

[4] McAfee, "McAfee Threats Report: Second Quarter 2013," retrieved on March 1, 2015 from http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q2-2013-summary.pdf.

[5] TrendMicro, "Android Malware: How Worried Should You Be?" retrieved on March 1, 2015 from http://blog.trendmicro.com/trendlabs-security-intelligence/android-malware-how-worried-should-you-be/

[6] F-Security, "MOBILE THREAT REPORT Q4 2012," http://www.f-secure.com/static/doc/labs_global/Research/Mobile%20Threat%20Report%20Q4%202012.pdf.

[7] Yerima, S. Y., Sezer, S., McWilliams, G., and Muttik, I., "A New Android Malware Detection Approach Using Bayesian Classification," the 27th IEEE International Conference on Advanced Information Networking and Applications (AINA), 2013.

[8] Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J., "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function," Journal of Computational Chemistry, 1998.

[9] Bridge, D., "Genetic Algorithms," retrieved on March 1, 2015 from http://www.cs.ucc.ie/~dgb/courses/tai/notes/handout12.pdf.

[10] Sarma, B. P., Li, N., Gates, C., Potharaju, R., Nita-Rotaru, C., and Molloy, I., "Android permissions: a perspective combining risks and benefits," The 17th ACM symposium on Access Control Models and Technologies, 2012.

[11] Wu, D.J., Mao, C. H., Wei, T. E., Lee, H. M. and Wu, K. P., "DroidMat: Android Malware Detection through Manifest and API Calls Tracing," The 7th Asia Joint Conference on Information Security, 2012.

[12] Di Cerbo, F., Girardello, A., Michahelles, F., and Voronkova, S., "Detection of malicious applications on android OS," Computational Forensics, 2011.

[13] Enck, W., Ongtang, M., & McDaniel, P., "On lightweight mobile phone application certification," The 16th ACM conference on Computer and communications security, 2009.

[14] Chiang, W. C., "Behavior Analysis of Mobile Malware Based on Information Leakage", master thesis of National Sun Yat-sen University, 2013

[15] Shabtai, A., Kanonov, U., and Elovici, Y., "Intrusion detection for mobile devices using the knowledge-based, temporal abstraction method," in Proc. Journal of Systems and Software, vol. 83, no. 8, 2010, pp. 1524-1537

[16] Cover, T. M., and Thomas, J. A., "Entropy, relative entropy and mutual information," Elements of Information Theory, 1991.

[17] Kouznetsov, P., "JAD Java Decompiler," retrieved on March 1 2015 from http://www.varaneckas.com/jad/.

[18] Lin, J. M., "Detecting Mobile Application Malicious Behavior Based on Taint Propagation", master thesis of National Sun Yat-sen University, 2013

[19] Aafer, Y., Du, W., and Yin, H., "DroidAPIMiner: Mining API-level features for robust malware detection in android," Security and Privacy in Communication Networks, 2013.

[20] Blasing, T., Batyuk, L., Schmidt, A. D., Camtepe, S. A., and Albayrak, S., "An android application sandbox system for suspicious software detection," The Fifth international IEEE conference on Malicious and Unwanted Software (MALWARE), 2010.

[21] McAfee Lab , 2012, "FakeInstaller' Leads the Attack on Android Phones," retrieved on March 1, 2015 from https://blogs.mcafee.com/mcafee-labs/fakeinstaller-leads-the-attack-on-android-phones

[22] Richardson, L., retrieved on March 1, 2015 from "BeautifulSoup," http://www.crummy.com/software/BeautifulSoup/.

[23] Neumann, M., "Mechanize," http://mechanize.rubyforge.org/.

[24] Zhou, Y. and Jiang, X., Android Malware Genome Project, retrieved March 1, 2015 from http://www.malgenomeproject.org/.

[25] Adrienne Porter Felt, Kate Greenwood, and David Wagner, "The effectiveness of application permissions," The Second USENIX Conference on Web Application Development, 2011.

[26] Wei, X., Gomez, L., Neamtiu, I., and Faloutsos, M., "Permission evolution in the android ecosystem," The 28th ACM Annual Computer Security Applications Conference. ACM, 2012.

[27] Grace, M. C., Zhou, Y., Wang, Z., and Jiang, X, "Systematic Detection of Capability Leaks in Stock Android Smartphones," The Annual Network & Distributed System Security Symposium (NDSS), 2012.

[28] Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B. G., Cox, L. P., and Sheth, A. N., "TaintDroid: an information flow tracking system for real-time privacy monitoring on smartphones," The 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI'10), 2010.

 ISBN: 978-1-61208-398-8

Lightbulb: A Toolkit for Analysis of Security Policy Interactions

Derrick Kong, David Mandelberg, Andrei Lapets, Ronald Watro, Daniel Smith, and Matthew Runkle
Raytheon BBN Technologies
Cambridge MA USA
email: {dkong, dmandelberg, alapets, rwatro, dsmith, mrunkle}@bbn.com

***Abstract*— Lightbulb is a toolkit for analysis of the combined impact of a set of diverse security policies. It is designed to securely access and collect the security policy configuration data from the hosts, routers, and firewalls that comprise a network enclave. Lightbulb loads the collected security configuration data into a modeling tool and allows system administrators to run queries against the model with the intent to verify desired security properties of the composite system. If a policy query fails, the user is given a specific instance of the policy violation that can be investigated and resolved. The overall toolkit provides an extensible framework for rigorous verification of security policies of network devices.**

Keywords-cyber security; security policy; network security policy; access control; logic programming; formal verification.

I. INTRODUCTION

Security configuration management has been a problematic issue ever since security devices have existed. In modern, heterogeneous networks, misconfigurations are not just occasional nuisances, but common problems that can lead to serious security breaches. Current work in the field has shown that rigorous verification of security policies is possible, but published research [1][2] has generally been limited to particular aspects of either policy or security configurations. The next logical step is to apply these principles across a broader spectrum of policies and security appliances and to compose multiple policies into a coherent system specification.

The primary challenge for building a coherent system-wide tool for managing security polices is that there can be dozens to hundreds of heterogeneous configuration files residing on devices in a typical enterprise network that will have an impact on some aspect of security. Without an easy way to collect, organize and provide end-to-end analysis, administrators must look at configurations in isolation or in small groups to verify that desired policies are being enforced.

This paper presents Lightbulb, an integrated toolkit of components that support rigorous automated security verification of a variety of network devices and clients. These components have been designed to fit within a general framework; individual components handle tasks, such as ingesting security policy specifications or query inputs (expressed in a Domain-Specific Language (DSL)), controlling and extracting the security-relevant configuration files from components located in a managed network and converting them to an intermediate form and, finally, performing rigorous verification of security policies against the configuration data. Models are built using Prolog within the Ciao logic programming environment and employ features such as constraint programming.

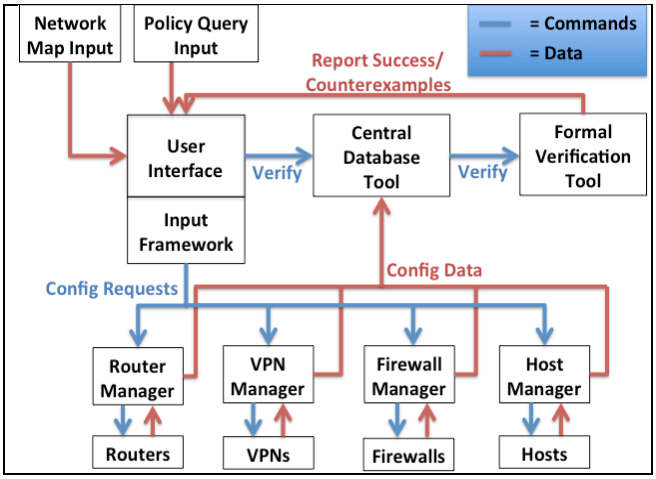


Figure 1. Toolkit architecture showing detailed data flows and interfaces with external inputs and devices to be managed.

The paper is organized as follows. Section 2 describes the general approach of the Lightbulb system. Section 3 provides an overview of the system user interface. Section 4 documents the currently supported network devices. Section 5 covers sample use cases and Section 6 covers currently supported queries. Related research efforts are described in Section 7 and the paper concludes with a summary and recommendations in Section 8.

II. APPROACH

A toolkit architecture diagram for the toolkit is shown in Fig. 1. Starting at the top left, a user provides inputs such as a network map and a security policy verification question via the User Interface. The Input Framework reads the map and uses it to identify the relevant devices in the network. It passes the results down to the Capability Manager Tools, each of which interfaces with one or more specific devices. An individual Capability Manager possesses the specific data needed to locate and interpret the security relevant configuration files from particular classes of devices; for example, the Router Manager is programmed with information on routers, including make, model and other relevant differences. (Support for basic Cisco devices covers a large portion of the install base; see Section VI for notes about extensibility.) A network device with more than one capability (for example, a firewall which is also a router) will report to more than one Capability Manager.

 ISBN: 978-1-61208-398-8

This architecture is easily extensible in that new Capability Managers can be added to the overall system and new device configurations can be programmed into existing Managers. Also, Capability Managers can easily be combined or split up as the Managers can be run on separate processors or the same processor. Our current architecture provides examples for the most common devices that should be present in a typical enterprise network. Extension and modification of these examples is a relatively straightforward process, allowing a local administrator to customize models for their devices, although a radically different architecture or device would require some significant work in order to translate its configuration and capability into our internal representation.

Once the configuration data have been obtained, the Capability Managers pass them up to the Central Database Tool, which stores them. The Database Tool also is responsible for processing the configuration files, including filtering and conversion to our Prolog representation for input to the Verification Tool.

Finally, the Formal Verification Tool takes in all the processed input that has been loaded into the Central Database, combines it with the input data from the User Interface, and performs the policy query check. The results of the check (either success or a listing of counterexamples) are sent back to the User Interface for display.

The formal verification module uses Ciao, a particular implementation of Prolog (with some extensions). Ciao is a modern tool for logic programming that supports strong modularization, which was often lacking in earlier logic programming systems. Ciao also includes support for constraint logic programming, which is very helpful when dealing with network packet data, and also tabling, memoization, and higher-order functions.

Logic programming is a very natural approach for modeling security policies, as both make use of the "negation as failure" concept [3]. In SELinux policy, for example, the active access vector rules are just the ALLOW rules, which permit certain actions. The lack of an ALLOW rule for an action means that that action is blocked. In our Prolog models for SELinux policies, the ALLOW rules become first order facts, and the Prolog engine can implicitly interpret the absence of a fact in a model as the assertion of the fact's negation, which matches SELinux semantics. There are indeed NEVERALLOW rules in SELinux, but they are passive specification rules, to be checked against the ALLOW rules after policy compilation. This approach to modeling SELinux policy using Prolog was first popularized by Scott Stoller and his student at SUNY Stony Brook [4].

If desired, a user can swap out the default formal verification module in the Lightbulb framework and replace it with a completely different engine or processor. This might be done if a particular security policy query requires a different logic or handling process. For example, the Accumulo query on prohibited label combinations employs a Python engine to handle the data processing as opposed to the Ciao engine.

III. User Interface Overview

The Lightbulb User Interface (UI) is designed to provide a single, convenient input and control interface for users, while still allowing developers an easy way of modifying existing functionality or adding additional modules to the toolkit. The UI provides the input and output mechanisms for the network map and policy queries; it works in close concert with a backend that handles operations for retrieving and managing configurations from networked devices.

The UI is based on open-web technologies. It makes use of three main frameworks:

- Flask, a python-based web server that handles processing
- Bootstrap, a frontend framework that provides templating and styling, and
- SigmaJS, a JavaScript graphing library used to draw network maps.

The UI provides a human interface to the Input Framework and negotiates communication with the Formal Verification Tool.

When Lightbulb is started, the user is presented with the configuration page, starting with the Network Map (Fig. 2). If this is the first run, the user is given the option of starting with a blank pane or to input an existing UML map; otherwise the last edited version of the Network Map is presented. At this stage, the user is free to alter the map to reflect the current network architecture. Lightbulb currently only read in maps files that use a UML deployment diagram representation.

The second configuration step allows the user to specify device credentials, including SSH-enabled, telnet-enabled, and Accumulo (a highly scalable database) node devices. Devices with common passwords can be supported as well as devices needing additional authentication (such as enable mode on Cisco devices).

In the final step of the configuration wizard, the Capability Managers are instructed to fetch device security policies. While the system contacts each device, a visual overlay is provided to indicate the status of the fetch process. Each device security configuration is retrieved and converted to an internal representation (expressed in Prolog) in parallel. The user interface periodically polls the server for updates, changing the status as fetches complete (or time out).

A. Security Policy Query Interface

The query page is an interface to the Prolog modeling engine. In order to make each type of query easy to formulate, the page provides a preset structure and list of options

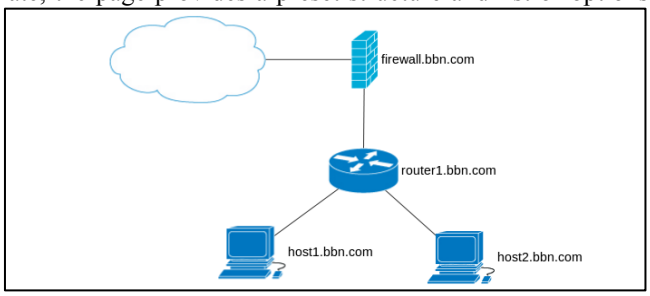


Figure 2. Network Configuration pane detail showing an example network.

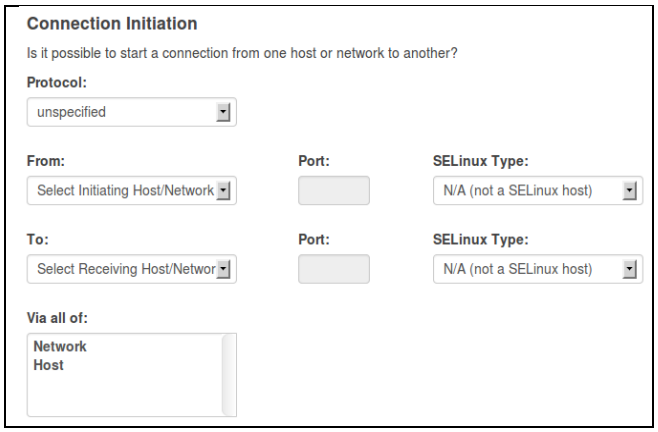


Figure 3. Query interface example. Shown is the selector for connection initiation queries. Values in the pulldown menus are automatically taken from the network map that is input by the user.

for each type of query. Submitting a query is as simple as selecting the appropriate options for each field and pressing submit. A small delay will occur as the system converts the information to the appropriate Prolog commands and submits them for processing. A simple "yes" or "no" response will return in typical Prolog fashion, along with counterexample data (in limited cases) if the answer is "no." An example of the query interface for reachability and connection queries is shown in Fig. 3 and a complete list of supported queries is given in Section 6.

IV. Currently Supported Devices

We currently support policy analysis for multiple types of devices: Iptables, Cisco routers/firewalls, SELinux, UNIX discretionary access control, and Accumulo hosts. Each of these analyses is discussed below.

A. Iptables

Iptables is an application program that allows a system administrator to configure the tables provided by the Linux kernel firewall and thus the firewall policy. We have completed a Prolog model that implements the iptables policy. Because iptables is so general, we are able to use our model as a basis for covering other packet polices, such as the Cisco IOS access list rules (discussed in the next section).

In constructing an iptables policy, the largest granularity item is called a table. Tables consist of one or more chains, where chains can be built-in or user-defined. Chains may contain multiple rules, where rules determine an action to take on packets.

There are four built-in tables: filter, NAT, raw, and mangle. Each built-in table contains a few built-in chains. Each rule in a chain contains a goal and a target. If the rule goal is matched, then processing continues onto the rules specified in the target. If the goal is not matched, then processes moves to the next rule. The default value in iptables is to accept if no rule applies, but this default can be changed by the policy. Lightbulb currently only considers the filter table for its policy analyses.

A rule in iptables may take a number of actions on a packet, including:

- Accept the packet
- Drop the packet
- Queue the packet for user interaction
- Return the packet to the calling chain.

An example of a very basic set of rules that Lightlbulb can parse and verify is shown below:

```
iptables -A INPUT  -j ACCEPT -p all -s 192.168.1.0/24 -i eth1
iptables -A OUTPUT -j ACCEPT -p all -d 192.168.1.0/24 -o eth1
```

These rules allow traffic to/from a specific subnet to pass through the firewall.

Our Prolog implementation of iptables policy uses nested lists to represent the chains and rules inside each table. We use recursive list traversal in order to ensure that the chains and rules are examined in order and that the first rule matching a particular connection (i.e., a hypothetical packet with particular source and destination addresses and ports) is selected for application.

B. Cisco Routers/Firewalls

The Cisco IOS support for router and firewall access control lists has evolved over the years and currently includes a wide array of options. In IOS, access control lists are numbered or named, and a numbered list can be applied to either the in-bound or out-bound traffic on an interface. Each access control list may contain explicit permit or deny rules, and the order of the rules is important, as the first match of a packet to a rule determines the status of the packet. There is also an implicit deny at the end of each access control list, so that if a packet matches no rule, it is rejected. Due to the presence of both explicit and implicit denial, we explicitly process Cisco access control lists to find the first match to a packet, rather than storing them as Prolog facts and allowing failure to define denial.

The general syntax for a CISCO extended rule is:

[permit/deny] protocol source destination parameters

The protocol can be one of IP, TCP, ICMP, and UDP. The exact syntax of a rule varies for each protocol. In general, the source and destination can be ranges of IP addresses, expressed using an IP address and a host mask; in this manner, "host 192.168.30.5" means the same as "192.168.30.5 0.0.0.0". Ports can be specified either numerically or by using names for the well-known ports.

For the current system, we have chosen to model the Cisco extended rule set, an example of which is verifiable by Lightbulb is shown here:

```
access-list 101 permit tcp any host 192.168.35.1 range 20 21
access-list 101 permit tcp host 192.168.30.5 host 192.168.35.1
   eq telnet
access-list 102 permit tcp host 192.168.35.1 any
access-list 102 permit tcp host 192.168.35.1 eq 20 any gt 1023
access-list 102 deny udp any

interface Ethernet0
access-group 101 out
access-group 102 in
```

 ISBN: 978-1-61208-398-8

C. SELinux

As described in Section 2 above, we start with the Prolog approach found in previous work [4] to build a model of the security policy for SELinux hosts. We include allow rules, type transitions, conditional rules and SELinux policy booleans in the model. In addition, we link the SELinux policy to the network activity by explicit tracking of both the assignment of port numbers to types and then enabled allow access rules for sockets associated with these port numbers. Under SELinux, the use of a network port number in TCP or UDP traffic communications is limited by policy and our tool analyzes these limitations when it performs checks on connections.

D. Unix Discretional Access Control

We include a Prolog model for the well-known discretionary access control in conventional Unix, based on settings for the user that owns the object, the group that the object is in, and all others. This was implemented without use of the Prolog cut operation, which is commonly used to control backtracking, but in this case would limit the usability of the definition. Instead of cuts, the group and other rules contain explicit hypotheses that express the proper order handling of user/group/world permissions.

Note that the user interface does not currently support preformed queries for discretionary access control, but it is accessible via direct Prolog input.

E. Accumulo

Apache Accumulo is a scalable data store based on Google's BigTable model. One of its enhancements beyond BigTable is the implementation of cell-level access controls which allow data cells at different security levels to be stored in the same table.

Accumulo clusters may be composed of a variety of complex network configurations and storage topologies; Lightbulb enables seamless reasoning over Accumulo cluster security by modeling policies at both the network and database configuration levels. While Accumulo may not be as common as other data storage technologies like MySQL, it uses access control permissions similar to those found in other databases and combines them with new types of security controls. As a result, the models and queries developed for Lightbulb's Accumulo support can be ported to support other database systems.

Like many common databases, an Accumulo cluster does not have a single well-defined policy file for the entire system. Instead, its security properties are defined by four separate configurations: the configuration files, user authentication, database access control permissions, and data cell visibilities.

Of these configuration types, the configuration files are of comparatively minor importance. The configuration files primarily specify the network configuration of the Accumulo cluster and the user authentication information required to provide management oversight of the system. They can be used to derive the network topology of the Accumulo cluster given direct access to an Accumulo master server, but they do not contain information on Accumulo user accounts, access controls, or visibilities. The other configuration types must be obtained by interacting directly with the Accumulo database.

The Accumulo configuration structure described above presents a challenge to the Lightbulb data ingest model. To interface with Accumulo, a new access module was developed to extract configuration files and interact directly with the Accumulo database to extract the security properties. Because each cell in an Accumulo database has its own visibility, it is impractical to recover all of the relevant security data. Lightbulb captures only the information required to construct a Prolog model of the basic system and access control configuration and does not attempt to recover the visibility of individual cells. As a result, some security queries are executed using the Prolog system model while queries requiring access to data cell visibility are executed using a Python query engine and run against the live Accumulo cluster.

V. USE CASES

In this section, we present a set of typical use cases in which a network administrator might desire to verify security policies. These use cases were used to derive the set of supported queries that are listed in the next section.

A. Typical Enterprise Network

The first use case is a typical enterprise enclave with two independent connections to the Internet, a DMZ zone, and an internal backbone that spans multiple subnets. A set of standard security policies applicable to many typical corporate and university networks is relevant here, such as the provision of particular public services in the DMZ to the outside world, but limited or no exposure of services and hosts inside the inner firewalls, except in particular cases to the DMZ (such as an internal database accessed by a DMZ web service). Verification of firewall, routing, and service access rules would be relevant in this case.

B. Multiple Enclave Enterprise

A variant of the first use case is one in which an enterprise consists of multiple independent enclaves separated across the Internet. In addition to the questions in the previous section, relevant policy questions include enclave-to-enclave communication configurations, such as verification that VPN traffic between enclaves is encrypted and is being correctly routed through the designated endpoints.

C. Combined Network and SELinux

The next use case is shown in Fig. 4. The policy question illustrated in the figure is whether it is possible for data to flow from Host A to Host Z. The key concept here is the need to combine security policies; in this case, the access control policies on firewalls 1 and 2 must be combined with the security policies on the host systems, which are assumed to be SELinux. The blow-up of Host C illustrates the issue that data may transition in type as it flows across an SELinux host.

We note that enforcement of SELinux types upon data transmitted across a network is generally not enabled. To

 ISBN: 978-1-61208-398-8

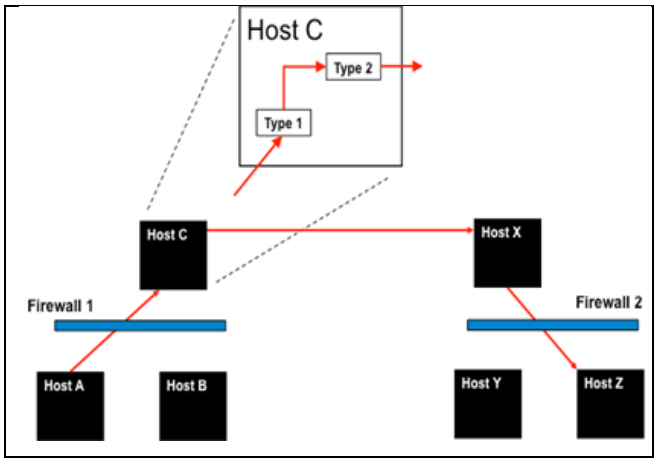


Figure 4. Basic combination of host (SELinux) and network policies.

do so would require all network hosts to be SELinux enabled or for mechanisms such as IPsec tunneling (with bindings established between Security Associations and types) between all relevant hosts to be enabled. While the latter is a more interesting case in terms of security policy, it is much less common in practice. Thus, we do not address this latter case in this version of the toolkit. Also, we assume that the Tresys tools are available to answer security policy questions that are completely self-contained on a single host.

Relevant queries then include basic routing questions as well as whether information flow can occur from one domain on an initial host to another domain on a destination host. For purposes of this work, we assume that identically labeled types on different hosts are equivalent; this is not enforced on general networks, but is often assumed to be so in practice. Type transition rules, as well as roles, are also relevant to policy queries, but require more advanced handling; Lightbulb currently does not support analyses using such transitions, but could do so with some additional development support.

D. Accumulo

In addition to the network connectivity and routing related queries described previously, a network administrator must be able to assess the security properties of the Accumulo database itself in addition to those of the hosts it runs on. Network related queries are still relevant for an Accumulo cluster, but new queries are also required to validate its user account, table access control, and cell visibility properties.

An administrator may need to determine whether Accumulo users have certain permissions (read, write, etc.) on tables in the database or hold other system-wide administrative rights, a type of query equally applicable to other databases. In addition to this general database permissions query, the data cell visibilities particular to Accumulo demand two unique queries.

The first addresses the need to determine if an Accumulo user account has the proper authorizations to view a cell with a particular visibility. The result allows an administrator to confirm that a user account's access to system data is not overly restricted by unintentional consequences of the applied security configuration.

The second type of query allows an administrator to determine if restricted cell data is leaking between users with exclusive authorizations due to cell-level misconfigurations. Accumulo's data cell visibilities allow the storage of data at different security levels within the same table; if cells are inserted with misconfigured combinations of security labels (e.g., the visibility is restricted to users holding SECRET or PUBLIC authorizations), restricted data can become visible to unintended users. This invalid visibility query allows an administrator to quickly identify every data cell violation in the system given a policy that describes the permitted relationships of security labels.

VI. Currently Supported Queries and Extensibility

The set of preformatted policy queries supported within Lightbulb is shown in Table 1. These policies were chosen as exemplars of each class of query; future work will extend these to related domains such as confidentiality, authorizations, etc. In addition to these, the Lightbulb interface allows an expert user to formulate an arbitrary query in Prolog that is passed directly to the Formal Verification Tool; no checks or constraints are applied to this query, so only expert users should attempt to use this option.

A user can create additional preformatted policy queries by formulating a native Prolog query and following the existing templates presented in the user interface to provide arguments. This process requires Javascript programming knowledge and understanding of the Prolog modules, which will require in-depth knowledge in those areas.

Lightbulb currently supports SELinux hosts, basic Cisco firewall/routers, hosts running IPtables, and Accumulo clusters. Lightbulb also provides a generic temple (in Python) that a user can adapt to read the configuration files from other similar device types. However, a device using a com-

TABLE 1: Currently Supported Preformatted Queries

Query Type	Query	Parameters
Reachability	Does a path exist between these hosts or networks?	From (host/network), To (host/network)
Connection	Is it possible to start a connection from one host or network to another?	Protocol, From (host/network), From (port number), To (host/network), To (port number), Via (hosts/networks)
SELinux	Is it possible for data of a one type on a SELinux host to transition to data of another type on another SELinux host?	Originating host, Originating type, Destination Host, Destination Type
Accumulo Permissions	Do users have a specified permission on particular tables?	Host, Users, Tables, Permission (read/write)
Accumulo Visibility	Are invalid visibilities present in the database?	Host, Tables, System Labels, Allowed Label Coexistence
VPN	Is it possible to start a connection from one host to another, going through VPNs?	Protocol, From (host/network), From (port number), Originating VPN, Receiving VPN, To (host/network/port)
Database (MySQL)	Is a user connecting from a host authorized with a particular permission on a particular table?	User, Host, Permission, Table

 ISBN: 978-1-61208-398-8

pletely new paradigm of storing and reading configuration data may require a new access method not currently supported by the existing work.

VII. Related Work

Early concepts related to the current Lightbulb toolkit where developed at BBN under the Cyber Command System (CCS) project in the DARPA Information Assurance program [5].

The applicability of logic programming for security policy modeling has been noted and exploited in numerous papers [4][6][7]. The early research of this type used the original Prolog language [8] but the later work, including this paper, employ more general logic programming techniques such as tabling and data constraints. Bounded model checking is another approach to the policy analysis problem, using tools such as Alloy [9] and Margrave [10]. The model checking approaches can excel at analyzing changes in security policies.

Previous work has addressed information flow polices in networks and in SELinux [1][2]. The current paper builds off that work by creating a single model that includes the policy requirements from all the components in an enclave.

Accumulo is a relatively recent system and as such, most existing research for it has been in developing architectures and analyzing performance; a few analyses of security have been performed [11], but not in conjunction with other systems.

Finally, there is much current research on Software Defined Networks (SDNs), where specific custom-built network hardware such as routers or firewalls are being replaced with generalized all-purpose network appliance with the power to dramatically redesign a network just by accepted new configuration data [12][13]. Creating assurance in the configuration settings of an SDN is a vitally important challenge that could be supported by future extensions of the Lightbulb toolkit.

VIII. Summary and Conclusions

This paper has presented the Lightbulb, an integrated set of toolkit components for networked system security analysis. Lightbulb allows a network administrator to collect and analyze the various access control and security policies that exist inside a network or collection of networks. Lightbulb includes a user interface tool for network system definition and also automated support for extracting policy data from various configuration files. A query tool is provided with templates for common policy statements. A user can submit a security assertion as a query and receive a verification that the assertion holds for the policy, or a counterexample to the assertion that the user can examine to refine the policy.

We plan to continue to develop the Lightbulb toolkit. Our highest priorities are to expand the set of supported queries and to construct models for new devices from the realm of Software Defined Networking.

References

[1] J. D. Guttman, and A. L. Herzog, “Rigorous automated network security management,” International Journal for Information Security, vol. 3, no. 3, pp. 29-48, 2004.

[2] J. D. Guttman, A. L. Herzog, J. D. Ramsdell, and C. W. Skorupka, “Verifiyng information-flow goals in Security-Enhanced Linux,” Journal of Computer Security, vol. 13, no. 1, pp. 115-134, 2005.

[3] K. L. Clark, “Negation as failure,” in Logic and Data Bases, H. Gallaire and J. Minker, Eds., Springer-Verlag, 1978, pp. 293–322, doi:10.1007/978-1-4684-3384-5_11.

[4] B. Sarna-Starosta, and S. D. Stoller, “Policy analysis for Security-Enhanced Linux,” In WITS’04: Workshop on Issues in the Theory of Security, 2004. Available at http://www.cs.sunysb.edu/~stoller/WITS2004.html [retrieved: February 2015]

[5] D. F. Vukelich, D. Levin, and J. Lowry, “Architecture for cyber command and control: experiences and future directions,” DARPA information survivability conference and exposition, vol. 1, DARPA Information Survivability Conference and Exposition (DISCEX II'01), Volume I, 2001, pp. 155-164.

[6] B. Hicks, S. Rueda, L. St.Clair, T. Jaeger, and P. McDaniel, “A logical specification and analysis for SELinux MLS policy,” Proceedings of the 12th ACM Symposium on Access Control Models and Technologies, June 2007, pp. 91-100.

[7] L. A. Wahsheh, D. Conte de Leon, and J. Alves-Foss, “Formal verification and visualization of security policies,” Journal of Computers, vol. 3, no. 6, June 2008, pp. 22-31.

[8] W. F. Clocksin, and C. S. Mellish, Programming in PROLOG, Springer-Verlag, 1981.

[9] D. Jackson, Software Abstractions, revised edition, Logic, Language, and Analysis, MIT Press, 2011.

[10] K. Fisler, S. Krishnamurthi, L. A. Meyerovich, and M. C. Tschantz, “Verification and change-impact analysis of access-control policies,” Proceeding ICSE '05 Proceedings of the 27th International Conference on Software Engineering, ACM New York, NY, 2005, pp. 196-205.

[11] M. Allen, “Past and Future Threats: Encryption and security in Accumulo.” Presentation at Accumulo Summit, June 2014.

[12] G. Lauer, R. Irwin, C. Kappler, and I. Nishioka, “Distributed resource control using shadowed subgraphs,” ACM Conference on Networking Experiments and Technologies (CoN-EXT), 2013, pp. 43-48, ISBN: 978-1-4503-2101-3.

[13] T. Nelson, A. D. Ferguson, M. J. G. Scheer, and S. Krishnamurthi, “Tierless programming and reasoning for software-defined networks,” NSDI'14 Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation, April 2014, pp. 519-531, ISBN: 978-1-931971-09-6

 ISBN: 978-1-61208-398-8

Multi-channel Secure Interconnection Design for Hybrid Clouds

Mauro Storch, César A. F. De Rose, Avelino F. Zorzo and Régio Michelin
Computer Science School – Pontifical Catholic University of Rio Grande do Sul
Email: {mauro.storch, regio.michelin}@acad.pucrs.br, {cesar.derose, avelino.zorzo}@pucrs.br

***Abstract*—Since its conception, Cloud Computing elasticity behavior offers a smart option of extension to companies' infrastructure. This extension is also used to create Hybrid Cloud Computing (HCC) environments through connecting private and public cloud instances. The connection channel between the company's private cloud and the public cloud provider is often encrypted due to security reasons. This paper proposes a multi-channel interconnection design for computing services of a hybrid cloud, targeting companies' security profiles. The design prioritizes vital channels, i.e., Quality of Services guarantees, and improves security by applying a different cryptographic cipher to each given channel. Also, our tests using OpenVPN as the channel player show improved communication up to 5 times when splitting the workload into multiple connections.**

***Keywords*—Hybird Cloud Computing; Communication Security; Distributed Systems; Network Communications; Communication Performance; Quality of Services.**

I. INTRODUCTION

HCC is becoming a scalable and economic solution to increase an in-house infrastructure. This architecture is a composition of public and private cloud models which support computational resources as services. The scalable skill is supported by the well-known feature called Cloud Computing Elasticity [1]. This feature allows companies to rent computational resources on-demand, from Virtual Machines to Storage area. Once a local environment is designed as a private cloud, it applies the Cloud principles and increases computational management quality, reduces the project budget through infrastructure sharing, and also saves power consumption by virtual machine consolidation.

Recently, VMWare vCenter started to offer a hybrid concept, called vCloud Hybrid Service [2], where companies move its virtual infrastructure to a public cloud in a secure and transparent way. However, there are some limitations in terms of heterogeneity in the proposed model. The VMWare's product only supports its own virtualization technology and the extension should be made to their public cloud instance.

On the other hand, Amazon Web Services (AWS), RackSpace, HP Cloud, among others, offer the majority of Cloud services on Internet [3]–[5]. Features such as global zoning design and the competitive price make them attractive for large, medium and even more for start-up companies. Through the global zoning design, users can easily deploy their applications around the world to reach more clients or users, considering short communication paths, local data processing, mobility and so on. Besides, each Cloud provider is priced on demand, the more expensive the resources (electricity, taxes, location, etc.) the higher the price per service unit is (month, hour, CPU cycles, bandwidth usage, etc). A company can arrange its outlays by balancing the elasticity distribution over a public cloud instance, considering aspects such as on-demand economy, availability, scalability, etc. [1].

Based on this scenario, one may consider an environment extension of a company infrastructure to some Cloud provider under the hybrid cloud model. Although a local environment is under secure and controlled rules, companies normally adopt a single communication channel through the Internet between the public and its local private cloud instance. This channel is used for data exchanging of applications, tasks synchronization, monitoring and so on.

Once this channel is created over the Internet, it is susceptible to different types of attacks such as eavesdropping, man-in-the-middle, data modification, and so on [6]. In this case, the application's data or even its integrity could be compromised. The most common communication setup is applying some *ad hoc* encryption algorithm in order to keep the exchanged information secure against these attacks. In order to ensure the security of this information, a strong encryption algorithm may be necessary.

This paper proposes to create multiple communication channels between the public and private cloud, each one for the different types of information that are exchanged between the Clouds. The information sent through these channels has different levels of priority. So, in order to achieve this qualification, it would be reasonable to assign different priority levels to each channel.

By separating the communication profiles (managing, monitoring, database replication, application's data exchange and many others) into different channels with different encryption levels, it is possible to keep information safe and also achieve better performance through load balancing among the created channels. Also, data exchange in the same application could be spread over those channels, after identifying the required priority level [7].

This work considers the HCC concept following the NIST (National Institute of Standards and Technology) [8] point of view and explores a secure multi-channel communication design in this concept.

The paper is organized as follows. Section II presents some related works. Section III describes the Multi-channel secure design for hybrid clouds. Section IV presents the multi-channel design application and the issues to be considered. Section V describes the proof of concept, the experimental evaluation and the achieved results. Finally, Section VI shows the conclusions and future works.

II. RELATED WORK

Researches related to Cloud communication are focused on a Cloud interconnection using IP-VPN techniques.

 ISBN: 978-1-61208-398-8

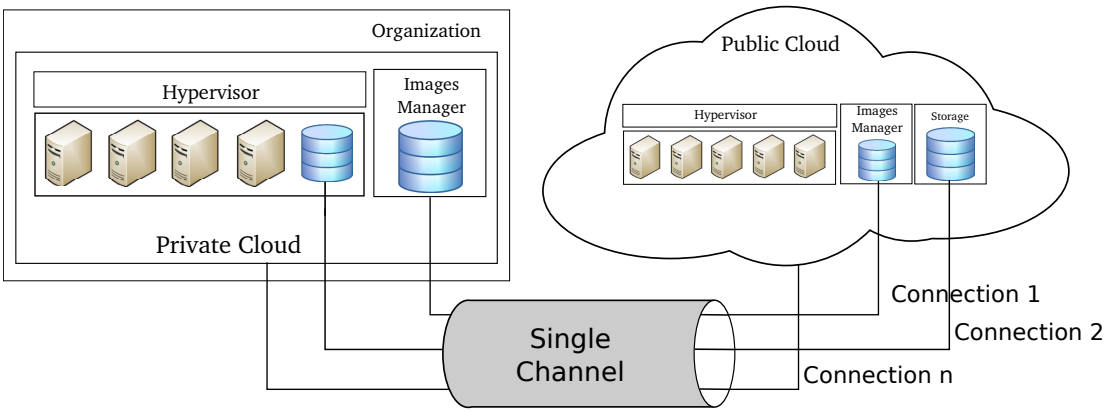


Fig. 1: Single Channel interconnection.

Ishimura [9] proposes a model in which a private company Intranet is connected to different cloud providers. This connection is established through IPSecVPN and is focused on the interconnection architecture (Full-Mesh or Hub). It does not consider the cryptography algorithm applied, neither the processing time involved in this operation.

Hata [10] defines a mechanism to connect different networks to a Cloud. Due to the virtual machine flexibility (it can be migrated between data centers depending on the load condition) he proposes a mechanism called Dynamic IP-VPN. He virtualizes the network, creating an architecture of Dynamic IP-VPN in order to enable users to control network components and equipment resources. In this architecture, he also describes a protocol that allows the user to reach the Cloud server. However, communication is still based on one tunnel connection per server's network interface.

Komu *et. al.* [11] specifies a model of secure communication for administrators, from the Infrastructure-as-a-Service (Iaas) point of view, using the Host Identity Protocol (HIP). The model isolates the multi-tenant network at the public cloud side and offers tunnelled access for external users. However, it does not consider security requirements, applying the same set of rules to the entire communication.

Wood [12] proposes the CloudNet architecture as a Cloud framework consisting of Cloud computing platforms linked with a network infrastructure based on Virtual Private Network (VPN) to provide seamless and secure connectivity between enterprise and Cloud data centers. This framework is optimized for supporting virtual machine live migration between geographically distributed data centers, but it still uses only one channel to migrate all information.

III. Multi-Channel Design

In order to build a hybrid cloud environment, an important component to be considered is the communication channel between the services on public and private clouds. This channel is commonly established through the Internet. The security requirements should be considered in agreement to company's data exposition rules.

Considering the interconnection among services of a private and a public cloud instances, a single channel using a VPN approach is normally adopted. All communication from each component, i.e., application's communication, database replication, service monitoring and management, are made through that single channel. Some improvements in WAN live migration of virtual machines also use VPN to provide network transparency after the migration [12]. Figure 1 shows the connections of some subsystems built through a single channel.

Usually, in common scenarios all data streams share a single VPN using the same physical path and configuration such as routes, network addresses, security strategies (considering authentication method and cryptography algorithms), and so on.

In a different perspective, a strategy that considers splitting isolated streams over different setups could bring new advantages in this scenario. For example, by qualifying the security priority of the communication for different parts of the hybrid cloud system, it is possible to rearrange the cryptography algorithms in order to achieve better results, i.e, by increasing either communication performance or security. For instance, during the application deployment, no sensitive data is transmitted. So, a non-encrypted channel could be used and the application could be verified on destination by comparing the generated MAC (Message Authentication Code). This verification ensures that the transferred data was not altered by any attacker during transmission. Moreover, some communication levels such as monitoring, management and billing, could also be encapsulated by lower encryption due to lack of business data on line.

Once the majority of inter-cloud communication channels are built over the Internet, the security ought to be considered an essential feature in this interconnection. Although, some IT managers penalize the security in order to acquire better performance during the communication. This trade-off relates to the fact that some encryption algorithms increase both the data size after encryption and the execution time in the processing phase (queuing application's communication). If the communication of services were split into several channels it would be possible to allocate different security levels for each

 ISBN: 978-1-61208-398-8

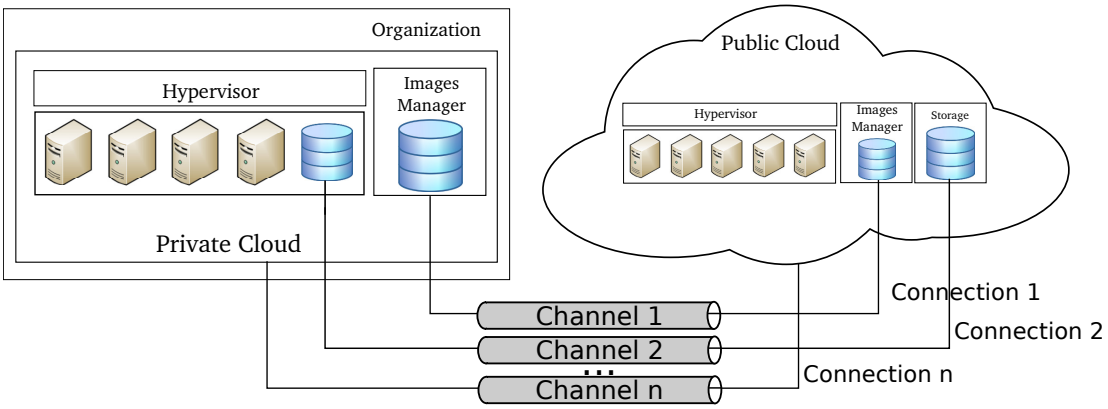


Fig. 2: Multi-Channel interconnection design considering hybrid cloud's communication split according to applications' security requirements, interconnection paths and isolation of communication's performance.

channel matching the security requirements of services (this security level differentiation will not be covered in this work). Figure 2 shows a channel per service connection.

An application defined as a three-layer model (data, logic and presentation) is a typical scenario in companies. The sensitive point would be the data and it should be protected in a more reliable communication channel when transferred outside the company's frontiers. This decision model considers that the logic and the presentation layers have already been deployed on public cloud resources and only data is present during the communication.

Although data is predominantly transferred, the security requirements are diverse for every company's application. Considering the three-layer stack for an application, some performance advantages would be achieved by splitting the communication channels and applying security techniques for the minimal requirements of each layer.

The multi-channel design would be implemented not only due to security issues. It is possible to identify at least two further benefits. A second benefit considers that a channel could be instantiated over different physical networks to balance or achieve better performance in communication. By using a network-layer tunnel protocol, such as IPSec [13], it is possible to attach the channel to a given logical or physical network, under a well-specified security model and drive the packet flow through a known route. In an upper level in the TCP/IP stack, it is possible to instantiate a transport-layer tunnel based on TCP or UDP protocols, by using tools such as the OpenVPN. The tunnel in that layer does not consider the layers below, which could be combined, expanding the security possibilities, or even replaced according to its network compatibilities.

A third benefit can be reached by adopting one channel per one or more connections, it is possible to schedule the total bandwidth by prioritizing a given channel. The balancing is provided by network techniques in each layer, such as filtering packets by origin and destination addresses, or even queueing content transferred such as backup or monitoring. On the one hand, from the hybrid cloud user point of view the communication would be transparent. On the other hand, from the administrative point of view, managers could prioritize critical communication related to services' life-cycle (monitoring, synchronization, replication, etc). Still, the production environment could be prioritized over software development resources, even into the same environment.

IV. Application Scenario

In order to instantiate the proposed interoperability design, this section presents a three-channel scenario considering a deployed application over a HCC environment.

The three channels are mapped in order to support the software communication between a private cloud instance and a public cloud provider.

The three channels are:

- **Application Synchronization:** Application will synchronize content and variables to keep its state;
- **Data Share:** Files and the database will place replicas among Cloud instances to provide fault tolerance, low access latencies, etc;
- **Management and Monitoring:** Access, authentication, monitoring, and billing processes are considered to manage public-side software.

The hybrid cloud scenario, instantiated for this environment, consists of an account in a public cloud provider and a private cloud instance. The public cloud account was created at AWS. It provides standard virtual machines to be connected to the private cloud ones. OpenStack [14] was used as private cloud manager. This tool is an open source initiative to support management of IaaS for common data centers. It is integrated with various hypervisors [15], such as KVM [16], XenServer [17], Hyper-V [18], and VMWare vSphere [19].

OpenStack was designed as a modular set of components that can interact with each other to create a single point of view of the entire data center's hardware. In this work only the Compute, Image and Networking modules are considered for managing the private cloud environment. These three modules support virtual machine creation and network configuration, but, in order to create the multi-channel design we use

 ISBN: 978-1-61208-398-8

OpenVPN [20]. The creation of channels can be a future feature to be added to cloud management tools to support the proposed model. Although OpenStack already supports VPN creation through IPSec, OpenVPN was used due to its firewall bypassing feature.

OpenVPN creates a single TCP or UDP connection in a given transport layer port, which could be set to be handled by the company's firewall. Once this connection is made, OpenVPN creates a virtual interface with a local-private IP address in both sides, i.e., client and server. From the user point of view, the interface is related to a local network connection. Every packet sent or received will be caught by OpenVPN, being encrypted and/or compressed, and delivered to the TCP/UDP connection.

Both cryptography and compression tasks are made in the operating system's user-space. So, the communication will use some CPU cycles both in private and public cloud. However, it does not have a significant impact on communication since OpenVPN creates a pipe during this process. In other words, as the packets are sent, they are one by one encrypted and delivered. Because the CPU is faster than the network communication there is no significant overhead. Yet, by using the Advanced Encryption Systems (AES) instructions provided by new processors, the CPU usage will be reduced as shown in Section V.

After the instantiation of the channels, based on this scenario, some features explore the benefits of the design as mentioned earlier. The first benefit is an increase in security due to the adoption of more than one cryptography key for the hybrid cloud interconnection. It is a straight forward achievement and it is supported through the size of the keys and the cryptography algorithm that will be adopted per channel. The second benefit related to paths and routes of connections were not implemented in this scenario, once this scenario was instantiated over a non-controlled network, the Internet.

The third benefit is the bandwidth scheduling, mentioned in Section III. It was implemented by prioritizing the channels according to its transferring profiles. Although the prioritization could be made without a specialized connection (using common TCP/UDP communications), it could be hard for IT managers to handle it on the fly. By setting up a rule in a given channel they could offer communication profiles for their users or applications transparently. These features are not commonly found on current instances of both public and private clouds, but they could be added to management tools as the following suggestions.

Although it is possible to manage both public and private clouds remotely through Application Programming Interfaces (API), the IT manager should be aware of every configuration in order to create the interconnection. The current OpenStack network module release can create a rich networking topology, including secure channels using VPN. However, there is no automation for setting up certificates and the cryptography level of each channel or even a prioritization behaviour. As a work in progress, we consider a contribution of those features by creating a proof-of-concept based on OpenVPN and OpenStack.

V. EXPERIMENTAL EVALUATION

In order to validate the multi-channel design, we show three evaluations focused on (i) demonstrating a performance increase by splitting communication in several channels; (ii) comparing the performance with different channel priorities and; (iii) empirically demonstrating an increased security provided by the multi-channel design that will be discussed during this section. The first test aims to show the adoption of a single VPN process (authentication, encryption, transferring, decryption) for each service in a hybrid cloud. By doing so, it will check if it is possible to reduce the overall communication time. The second test presents the results of adding prioritization of each channel. For the tests, Linux HTB [21] was applied to handle each level of the communication. This test is intended to show a bandwidth balance among the hybrid cloud's services.

A. Testbed description

For the tests, a hybrid cloud environment was configured combining a private cloud and a public cloud. The private cloud is composed by a machine powered by two Intel Xeon E7-2850 2.00 GHz (20 cores each, supporting hyper-threading) with the hardware virtualization flag enabled and using AES instructions, making cryptography phase up to five times faster in this scenario. Virtualization is provided by the Xen hypervisor [17] and managed by OpenStack (Havana 2013.2 release). OpenStack is responsible for the creation of virtual machines, local network configurations, storage allocation, and other aspects related to the management of the virtual resources. For the public cloud we use AWS, where a virtual machine was created under the IaaS model. This virtual machine is configured as a one core Intel Xeon CPU E5-2650 2.00GHz with 590MB RAM running Ubuntu Server 12.04 Amazon-image-based.

The three channels connecting the public cloud and the private cloud were implemented using the OpenVPN tool [20]. This tool is a standard open-source VPN player that offers TCP and UDP connections for data transport between peers under a client-server model. The server listens on a TCP/UDP port that is handled by both company's and public cloud's firewall. The transfer process supports cryptography and/or compression algorithms that are set during channel setup. For the tests in this paper, all the channels are configured using Advanced Encryption Standard (AES) [22]. Three encryption levels were evaluated during the experiments: AES128, AES192 and AES256. We also evaluate a channel without an encryption algorithm and the compression flag was disabled in all scenarios. In the public cloud's VM, a VPN service is instantiated in order to accept connections from the private cloud. There is one VPN service for each channel.

For performance evaluation, the communication is made over a TCP connection running in the User-Space domain. This approach enables tunnel creation using a TCP port that can bypass companies' firewall rules without exposing its network. Each connection is handled individually by the pair client-server. For new connections, a new server instance is created on another TCP port. Because the cryptography process of a tunnel runs in single-thread mode, by splitting

 ISBN: 978-1-61208-398-8

the communication into several tunnels, each one running on a different operating system process in parallel, we obtain a better performance of the cryptography process and consequently a reduction in the overall communication time.

The setup for QoS tests consider that several prioritization models could be applied for a given set of connections. In some use cases it is common that some connections need more bandwidth or even more priority than others. One can consider, for instance, a real-time application that needs to synchronize data frequently. Instead of turning off other connections to reduce the response time of the synchronization, it is possible to prioritize that channel to handle it without killing other service's communication. In the same way, in a scenario with a production and a test environment, the prioritization of the channel of production against the test environment could guarantee QoS aspects even in a hybrid cloud, without the need for multiple instances of a private cloud, one for each purpose.

The setup also considers that each channel has its own authentication certificates (generating a new cryptography key) and cryptography algorithm. The adoption of different encryption keys for the safe channels makes the HCC interconnection stronger, since an attacker needs to break all three cryptographic keys to intercept the entire information, instead of only one in the case of a single channel.

Since security improvements are straightforward, we have conducted experiments to validate the performance and QoS benefits of the multi-channel approach. Section V-B describes these evaluations and presents our preliminary results.

B. Evaluation and results

The first evaluation was conducted to showcase the performance of the multi-channel design. We first created pairs of client-server instances from two to ten. Payload was one GByte of data split equally among 2 to 10 channels along the horizontal axis. The *Transfer Time* line in the chart, Figure 3, indicates the execution time of the total data transfer. The *CPU Load* line indicates the aggregated CPU load in the client side considering all processes of the OpenVPN tool. In this case, only the tunnel was considered, discarding the application load time. In the chart, for example, 4% of CPU-load represents that the process took 4 seconds using the CPU for a 100 seconds communication. The CPU-load time is not necessarily a blocking operation, since packets have no more then 1500 bytes and are processed like in a queue. In other words, the CPU time has no significant impact during the communication. However, if the AES instructions were turned off, the CPU-load would be increased up to 20%, taking longer for packets to queue, adding more overhead for fast connections. In our scenario we considered all available resources, since AES instructions are commonly found in nowadays processors [23]. The cryptography algorithm applied to all channels was AES-256 and no compression was used. AES-128 and AES-192 were also tested, but no significant difference was found to be considered in the evaluation.

We can clearly see a reduction in transfer time as more channels are used. Although the CPU usage also increased, it is not significant, not exceeding 5% of the total CPU available in the machine with this technique. Nevertheless, communication time was significantly reduced, up to almost 5 times with 10 channels (from 532 to 108 seconds).

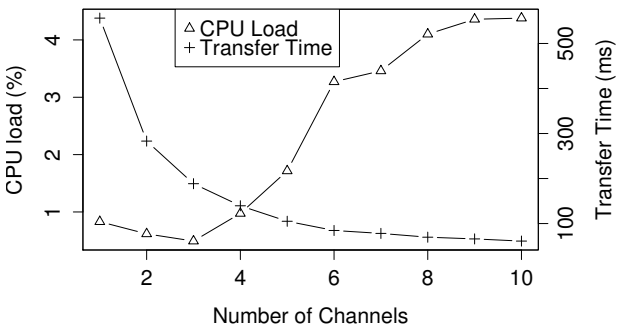


Fig. 3: Multi-Channel Splitting - Time x CPU load

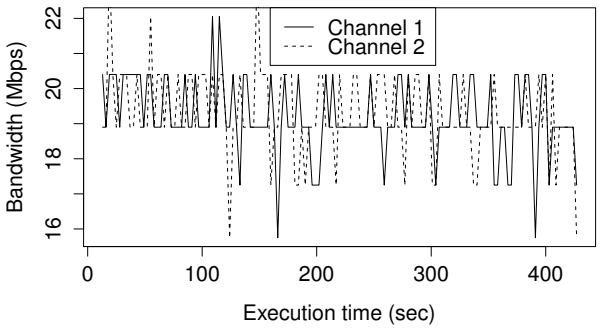


Fig. 4: Non-Priorized channel communication

The second evaluation which considered the QoS test was done to evaluate the effectiveness of prioritizing channels. In this case, we use a payload of one GByte per channel, sharing the same physical connection. Figure 4 shows the bandwidth of each channel and the resulting transfer time when both channels have the same priority concurring for the total system bandwidth (100 Mbits/s).

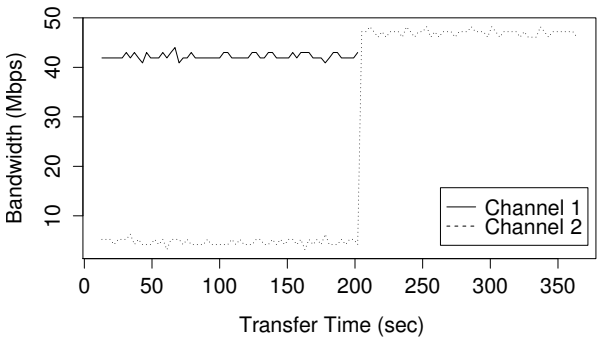


Fig. 5: Priorized channel communication

The possibility of prioritizing one of the channels results in better communication performance for sensitive data. This

can be clearly seen in Figure 5 where channel one has priority 0 and channel two priority 5 (lower values represent higher priorities). In this case, Channel 1 gets a much bigger share of the total system bandwidth resulting in a significant reduction of its transfer time (from approximately 420 seconds to 200 seconds). Channel 2, on the other hand, had a much smaller share of the bandwidth when concurring with channel 1, improving only after Channel 1 transfer was finished. This is an important feature of a multi-channel interconnection having several applications in real hybrid clouds use cases.

VI. Conclusions and Future Work

HCC plays an important role combining the advantages of private and public clouds. It allows managers to use local infrastructure in a flexible and efficient way and on demand rent services under a pay-as-you-go fashion from a public cloud provider, reducing acquisition and management costs with a more flexible environment.

The most common interconnection design in hybrid clouds uses a VPN to create a single channel between the private and the public cloud that consolidates all service flows. This channel is often encrypted due to security requirements of a specific traffic.

In this paper, we presented a multi-channel design that uses one channel per service. This allows a much more customizable design, with each channel being configured depending on the characteristics of the different flows. This increases the security and the performance of the communication with the public cloud. By applying different keys in each channel, for example, there is an increase in security with several keys to break in order to acquire information from all flows. Besides that, a stronger key/algorithm could be applied to sensitive data in one specific channel. Concerning the performance, the multi-channel design allows a prioritization model, where a channel with critic data could receive a greater bandwidth share to achieve better performance.

Our preliminary tests with the multi-channel design show also a reduced communication time when splitting the payload transferring in several channels by a factor of 5. The results were validated using OpenVPN tool, a TCP/UDP tunnel player which encrypts data before sending data in a user-space thread. The splitting improvement is related to cryptography parallelism provided by multiple threads. Nevertheless, the CPU load does not overpass 4.4%.

As future work, we will evaluate the impact of the multi-channel design when running a distributed application over a hybrid cloud. We are also interested in security overhead modelling in hybrid clouds.

Acknowledgment

The authors would like to thank CNPq (National Counsel of Technological and Scientific Development - Brazil), CAPES and FAPERGS for financial support.

References

[1] B. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *Fifth International Joint Conference on INC, IMS and IDC, 2009. NCM '09.*, Aug 2009, pp. 44–51.

[2] VMware, Inc., "VMware vCloud ® Hybrid Service ™ Service Description," http://www.vmware.com/files/pdf/vchs/vCloud-Hybrid-Service-Service-Description.pdf, retrieved: Mar., 2015.

[3] RackSpace US Inc., "RackSpace - The open cloud company," http://rackspace.com, accessed: Mar, 2015.

[4] Hewlett-Packard Development Company, "HP Public Cloud," http://hpcloud.com, accessed: Mar, 2015.

[5] Amazon, Inc., "Amazon AWS," http://aws.amazon.com/, accessed: Mar, 2015.

[6] S. Hansman and R. Hunt, "A taxonomy of network and computer attacks," *Computers & Security*, vol. 24, no. 1, pp. 31 – 43, 2005.

[7] P. Watson, "A multi-level security model for partitioning workflows over federated clouds," *Journal of Cloud Computing*, vol. 1, no. 1, pp. 1–15, 2012. [Online]. Available: http://dx.doi.org/10.1186/2192-113X-1-15

[8] P. Mell and T. Grance, "The NIST definition of cloud computing (draft)," *NIST special publication*, vol. 800, no. 145, p. 7, 2011.

[9] K. Ishimura, T. Tamura, S. Mizuno, H. Sato, and T. Motono, "Dynamic ip-vpn architecture with secure ipsec tunnels," in *2010 8th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT)*, June 2010, pp. 1–5.

[10] H. Hiroaki, Y. Kamizuru, A. Honda, T. Hashimoto, K. Shimizu, and H. Yao, "Dynamic ip-vpn architecture for cloud computing," in *2010 8th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT)*, June 2010, pp. 15–20.

[11] M. Komu, M. Sethi, R. Mallavarapu, H. Oirola, R. Khan, and S. Tarkoma, "Secure networking for virtual machines in the cloud," in *Cluster Computing Workshops (CLUSTER WORKSHOPS), 2012 IEEE International Conference on*, Sept 2012, pp. 88–96.

[12] T. Wood, K. K. Ramakrishnan, P. Shenoy, and J. van der Merwe, "CloudNet: Dynamic Pooling of Cloud Resources by Live WAN Migration of Virtual Machines," *SIGPLAN Not.*, vol. 46, no. 7, pp. 121–132, Mar. 2011. [Online]. Available: http://doi.acm.org/10.1145/2007477.1952699

[13] S. Kent and K. Seo, "Security Architecture for the Internet Protocol," RFC 4301 (Proposed Standard), Internet Engineering Task Force, Dec. 2005, updated by RFC 6040. [Online]. Available: http://www.ietf.org/rfc/rfc4301.txt

[14] OpenStack OpenSource Community, "OpenStack Project," http://www.openstack.org/, accessed: Mar, 2015.

[15] OpenStack OpenSource, Community, "OpenStack Hypervisors Support Matrix," https://wiki.openstack.org/wiki/HypervisorSupportMatrix, accessed: Mar, 2015.

[16] Qumranet, "KVM," http://www.linux-kvm.org/page/Main_Page accessed: Mar, 2015.

[17] P. Barham *et al.*, "Xen and the art of virtualization," *SIGOPS Oper. Syst. Rev.*, vol. 37, no. 5, pp. 164–177, Oct. 2003. [Online]. Available: http://doi.acm.org/10.1145/1165389.945462

[18] Microsoft, "Microsoft Hyper-V Server 2012 R2," http://www.microsoft.com/en-us/server-cloud/hyper-v-server, accessed: Mar, 2015.

[19] VMWare Inc., "VMWare vSphere Hypervisor," http://www.vmware.com/products/vsphere-hypervisor/, accessed: Mar, 2015.

[20] J. Yonan, "OpenVPN–an open source SSL VPN solution," https://openvpn.net/, accessed: Mar, 2015.

[21] M. Devera, "Hierarchical token bucket theory," *URL: http://luxik.cdi.cz/ďevik/qos/htb/manual/theory.htm*, accessed: Mar, 2015.

[22] N.-F. Standard, "Announcing the advanced encryption standard (aes)," *Federal Information Processing Standards Publication*, vol. 197, pp. 1–51, 2001.

[23] S. Gueron, "Intel advanced encryption standard (AES) instructions set," *Intel White Paper, Rev*, vol. 3, pp. 1–94, 2010.

 ISBN: 978-1-61208-398-8

A Generalized Approach to Predict the Availability of IPTV Services in Vehicular Networks Using an Analytical Model

Bernd E. Wolfinger [1)], Edgar E. Báez [2)], Nico R. Wilzek [3)]

[1), 3)] Department of Computer Science,
Telecommunications and Computer Networks
University of Hamburg
Hamburg, Germany
e-mail: wolfinger@informatik.uni-hamburg.de,
4wilzek@informatik.uni-hamburg.de

[2)] Superior School of Computing
National Polytechnic Institute ESCOM-IPN,
Mexico City, Mexico
e-mail: ebaeze0700@alumno.ipn.mx

***Abstract*— Currently, an increasing number of vehicles are getting equipped with components which offer the possibility of an Internet access with low expenditure. Therefore, entertainment services in VANETs are becoming more and more important. An interesting class of entertainment services comprises IP television (IPTV) services and, therefore, studies regarding the quality of experience (QoE) for IPTV in VANETs are becoming increasingly relevant. Such QoE analyses also constitute the main goal of this paper where we focus on QoE in terms of availability of IPTV services. Up to now, studies of IPTV service availability in VANETs have primarily been executed based on simulation models. In this paper, we make use of analytical models to predict availability of IPTV for VANET scenarios. For this purpose, we have elaborated an analytical model of rather low complexity which, nevertheless, is rather realistic as we will show by means of comprehensive validation studies. In addition, we propose a general proceeding which makes use of our analytical model and can be applied as a straight-forward approach to predict the availability of IPTV services in a flexible and efficient manner. Case studies demonstrate how our analytical model can be applied by a provider of IPTV services, offered via VANETs, in order to satisfy QoE requirements regarding the service availability as given by the IPTV users.**

Keywords- Vehicular networks; IPTV; QoE; service availability; analytical model; validation.

I. INTRODUCTION

Current predictions for the car market claim that, in 2016, more than 80 % of all new cars sold will have access to the Internet (e.g., FOCUS Online [5]). Therefore, one can expect that the usage of Internet services by car passengers will become more and more wide-spread in the near future. Besides search-, information- and communication services also entertainment services (such as IPTV or Video-on-Demand) will probably play a significant role [3]. For that reason, quality assessment of Internet services with real-time requirements (as they are present, e.g., in IPTV services offered in vehicular ad-hoc networks – or VANETs for short) is getting increasingly important. Therefore, this topic is in the main focus of this paper.

Quality of service provisioning is relevant, in particular as it is experienced by the (human) end-users and thus it is denoted by Quality of Experience (QoE) [9]. In case of IPTV services, QoE on the one hand refers to the quality of the received audio/video stream as perceived by the end-user [15]. But, on the other hand, it also comprises the degree of availability with which the user is able to access the IPTV service [8]. As a measure of availability, we will take the probability that a desired TV channel can indeed be provided to the corresponding user though the bandwidth in the (access) network may be quite limited. Availability studies for IPTV services have been done in the past (by means of simulation models) in particular for DSL based access networks [7] as well as for WiMAX based access networks [1].

As currently no vehicular networks offering IPTV services are available to us for carrying out measurements, the only alternative for corresponding service availability studies is the use of models. To the best of our knowledge, up to now, only very few models exist which allow one to predict the availability of IPTV services in VANETs. Detailed simulation models have been elaborated and applied in case studies by Momeni et al. [10] [11]. Moreover, in recent past, first successful trials have been undertaken to predict IPTV availability in VANETs by means of analytical models, cf. Wolfinger et al. [16].

This paper now significantly extends the results of [16] as we carry out an in-depth validation of the analytical model and, as a major new contribution, it presents a generalized procedure which allows us to predict the IPTV availability in a straight-forward manner for very different traffic scenarios and network technologies. We also apply our procedure in various comprehensive case studies.

The paper is structured as follows: Section II will give a short overview on IPTV services offered via VANETs including the availability measures which we will apply. The analytical model used will be introduced in Section III followed, in Section IV, by a thorough validation of this model. A generalized procedure for a highly efficient usage of this model then is presented in Section V. Application of the generalized procedure will be illustrated in the case studies of Section VI. These studies also show how our model can support a provider of an IPTV service (offered via a VANET) in dimensioning and configuring a network

 ISBN: 978-1-61208-398-8

which satisfies the given QoE requirements of the IPTV subscribers.

II. IPTV SERVICES IN VANETS AND AVAILABILITY MEASURES FOR THEIR ASSESSMENT

A. *Provisioning of IPTV Services in VANETs*

Two main classes of vehicular networks are typically distinguished: networks supporting vehicle-to-vehicle (V2V) and those supporting vehicle-to-infrastructure (V2I) communication. For our studies, only V2I configurations are relevant because communication between vehicles is not of interest to us. V2I communication can be achieved in two variants which differ in the way how users in the vehicles can get access to the Internet: in the first variant (V1), the mobile station (e.g., a smart phone) could be communicating via a non-IP-based public mobile network and from there get access to the Internet. In the second variant (V2), the mobile station would access a dedicated road-side unit (RSU) via the base station (BS) / the access point (AP) of its local cell and from there get direct access to IP based routers (cf. proposal and prototype for so-called road-side backbone networks using RSUs to interconnect the Internet with the vehicles as described, e.g., by Krohn et al. [6]). In this paper, we assume that the IPTV services which we analyze are provided in networks in which Internet access is established according to variant V2. Different network technologies (such as WLAN, LTE, WiMAX) can be used in principle to achieve communication between the mobile stations (in the vehicles) and the base station resp. access point in the corresponding cell. From point of view of IPTV service, provisioning different network technologies in the access network can have a strong impact on the service quality because they will typically support very different data rates and lead to very different cell sizes.

In the vehicular networks which we investigate, the fact that ad-hoc networking is possible between vehicles is not really important for us. On the contrary, we are mainly interested in the delivery of IPTV services to the vehicles by means of vehicle to infrastructure (V2I) communications. Nevertheless, we argue that the IPTV service delivery studied in this paper does not only cover vehicular networks, but also VANETs and, accordingly, we use the formulation "IPTV Services in VANETs" throughout this paper.

If an IPTV service is offered in a network with V2I communication where Internet access is achieved by means of RSUs (as assumed in our studies) the basic network architecture will comprise the main components as depicted by Figure 1:

- the IPTV Head-end, where all the TV channels are available which can be demanded by the IPTV users,
- that part of the Internet which is used to make communication between the Head-end and the set of RSUs possible (this subsystem could be the IP based network of an ISP providing the IPTV service),
- the access network representing the infrastructure for communication between an RSU and the mobile stations within the cells for which RSU is responsible.

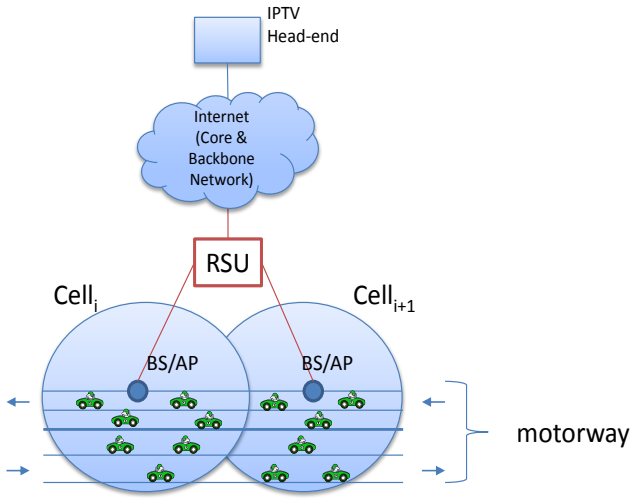


Figure 1. Basic architecture of an IPTV system for users of vehicular networks

Provisioning of IPTV services typically makes use of multicast (e.g., IP multicast [13]) leading to the advantage that a TV channel having been desired in an access network has only to be provided once by the corresponding RSU even in the case that the TV channel is currently watched by more than one user in this cell. A TV channel is no longer transmitted in a cell as soon as the last user watching this channel releases the channel (e.g., because he/she switches to another channel or is involved in a handover thus leaving the cell or the user may temporarily terminate usage of the IPTV service).

As a consequence of the limited data rate (bandwidth) of the cells representing the access networks it is, of course, possible that a TV channel newly desired by a user cannot be provided at that moment when the request for the channel is issued. This happens exactly in the case that the desired channel currently is not yet delivered in the cell AND the total transmission capacity available for IPTV is completely exhausted currently because of having to transmit other channels. If a request for channel delivery has to be denied, we say that the channel is "blocked" for the user and call this event a *"blocking (event)"*.

So, we see that studies of IPTV service availability in VANETs which are based on detailed models will require that the corresponding models reflect

- the bandwidth utilized for IPTV at any instant,
- the list of TV channels currently being multicast in the corresponding cell
- the behavior of the IPTV users in terms of the time instants at which TV channels are switched/changed and in terms of the id. (e.g., channel number) of the channel newly demanded.

Former investigations with respect to a realistic characterization of IPTV user behavior [1] [2] have shown that the popularity of TV channels can be approximated quite well by Zipf distributions [12].

 ISBN: 978-1-61208-398-8

In particular, the probability p_i that the *i*-th popular channel is requested is determined by the Zipf distribution as follows:

$$p_i = \frac{1/i^{\theta}}{\sum_{k=1}^{N} (1/k^{\theta})}$$

where *N* denotes the total n° of different channels offered, *k* is their rank and θ is the Zipf parameter reflecting the degree of popularity skew. A value of θ = 1.3 is realistic according to measurements of IPTV user behavior [1].

B. Measures for IPTV Availability

The following two reasons exist that an IPTV user will demand a TV channel within a cell:

(1) A *channel-switching event*: Here, the user will demand a new channel to which he currently switches to (e.g., because he is "zapping" through a sequence of channels at time durations of just a few seconds or after he terminates a "viewing phase" with duration of several minutes or even hours during which he has received and watched just a single TV channel).

(2) A *handover event*: Here, the car will change the cell and, as a consequence, the channel currently received by a user in this car will no longer be needed by him in the "old" cell left but it will be needed in the "new" cell reached now.

In both cases, blocking of the desired channel may occur. Thus, we distinguish:

– *switching-induced* or *switching-related blocking*, and

– *handover-induced* or *handover-related blocking*.

Therefore, three channel blocking probabilities are of interest to us:

– *Channel Blocking Probability* (CBP) referring to all blocking events

– *Switching-induced Blocking Probability* (SBP) referring only to blockings being a consequence of channel switching

– *Handover-induced Blocking Probability* (HBP) referring only to blockings being a consequence of handover events.

As it is usual, we can approximate the three probabilities by the relative frequencies of the corresponding blockings choosing an observation interval which is sufficiently large.

Let T = [t1, t2] denote the observation interval and |T|=t2-t1 its length.

Let further denote:

- #r(T): n° of all channel requests issued by all users in T
- #r_h(T): n° of all handover-related requests in T
- #r_s(T): n° of all switching-related requests in T
- #b(T): n° of all blocked requests (blockings) in T
- #b_h(T): n° of all handover-related blockings in T
- #b_s(T): n° of all switching-related blockings in T.

Based on these variables, we can now define the following channel blocking frequencies for the interval T:

- CBF(T) $\triangleq \frac{\#b(T)}{\#r(T)}$ denoting the *overall channel blocking frequency*
- HBF(T) $\triangleq \frac{\#b_h(T)}{\#r(T)}$ denoting the *relative frequency of handover-related blockings*
- SBF(T) $\triangleq \frac{\#b_s(T)}{\#r(T)}$ denoting the *relative frequency of switching-related blockings.*

Evidently,

$$\text{HBF(T)} + \text{SBF(T)} = \frac{\#b_h(T)}{\#r(T)} + \frac{\#b_s(T)}{\#r(T)} = \frac{\#b_h(T) + \#b_s(T)}{\#r(T)} = \frac{\#b(T)}{\#r(T)} = \text{CBF(T)}$$

and – as the relative frequency converges to the probability for an interval length |T| tending to infinity:

$$\text{CBP} = \lim_{|T|\to\infty} \text{CBF(T)}$$
$$\text{HBP} = \lim_{|T|\to\infty} \text{HBF(T)}$$
$$\text{SBP} = \lim_{|T|\to\infty} \text{SBF(T)}$$

which implies that also CBP = HBP + SBP holds.

Instead of CBP we can alternatively use

$$\text{CA} \triangleq 1 - \text{CBP}$$

denoting the overall *channel availability*.

III. An Analytical Model to Predict TV Channel Availability

In [16], an analytical model was elaborated which is the basis of this paper. This analytical model is used to determine CBP and it is able to take into account various traffic scenarios, access network technologies and IPTV service characteristics.

To present this model in this section and in the following sections we use the variables and model parameters as introduced in Table I.

The basic ideas underlying the analytical model are the following ones:

TABLE I. LIST OF PARAMETERS AND VARIABLES USED

	Variable/ parameter	Meaning
Traffic-related variables	k	number of lanes per direction
	v_i	speed of vehicles on lane L_i assumed to be constant for this lane (in [km/h])
	d_i	distance between adjacent vehicles on L_i assumed to be constant for this lane (in [m])
	$\bar{d}$	mean distance between adjacent vehicles (averaged over all lanes)
Cell-related variables	C_r	radius of cell (in [m])
	BW_c	bandwidth available for IPTV service in cell c
	N_c	number of IPTV users in cell c
IPTV-related variables	N	number of TV channels offered in total
	α	percentage of vehicles using IPTV
	p_i	probability that channel i is required (according to Zipf distribution with parameter θ)

(1) Calculate the probability that the system is in a state in which blocking can occur, also called a *"potential blocking state"*.
(2) Calculate the probability that a currently unavailable channel is demanded when the system is in a "potential blocking state".

Calculation of CBP in our analytical model is based on the following four steps:

- STEP 1: Determine the probabilities P_i that, for given N and N_c, exactly i different channels are needed to satisfy the channel requests of N_c users, if N different channels are offered. P_i can be estimated by the relative frequency f_i that N_c users require exactly i different channels, where f_i can be determined in a straight-forward manner by means of Monte-Carlo simulation [4] [14]. Throughout this paper, all of our Monte-Carlo experiments are repeated one million times and, therefore, the size of the sample to calculate f_i is 10^6.
- STEP 2: Assume a certain cell bandwidth BW_c available for IPTV and determine P^* as probability that N_c users require more than BW_c different TV channels. So, P^* denotes the probability that the system is in a "potential blocking state".
- STEP 3: Assume that an IPTV user will require a new channel (channel number determined according to Zipf distribution) and determine the probability that the number of the channel demanded is larger than BW_c , which happens with probability

$$\sum_{i > BW_c}^{N} p_i$$

- STEP 4: We determine the probability (*CBP*) that a newly requested channel cannot be delivered which happens with probability

$$CBP = P^* \bullet \sum_{i>BW_c}^{N} p_i ,$$

if we make the favorable assumption that, in case of a "potential blocking situation (state)", exactly those channels are transmitted in the corresponding cell, which are the BW_c most popular ones.

Remark: It should be noted that, astonishingly, the (favorable) assumption that "if the system is in a potential blocking state then just the most popular channels are transmitted" is quite realistic indeed. This has been observed by us in simulation experiments based on detailed models of IPTV services in VANETs (cf. simulation models described in [11] and also used during our model validation in Section IV). □

Figure 2 illustrates STEP 1 and STEP 2, by way of example, if we assume N = 50, BW_c = 30, N_c = 150. This figure depicts the histogram for P_i , $i \in \{1, 2, \ldots, 50\}$.

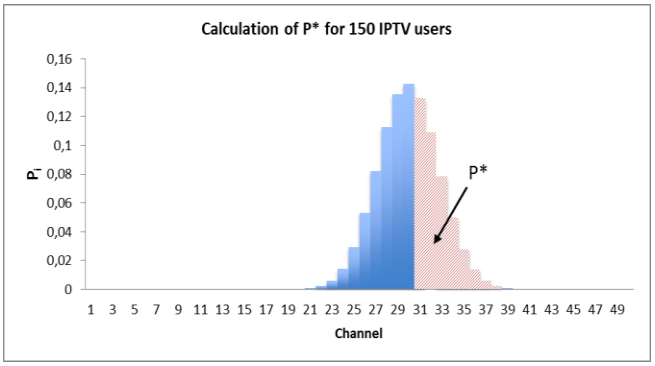


Figure 2. Determination of P^* for N = 50 and BW_c = 30 according to STEPs 1 and 2 of our calculation algorithm for the analytical model (N_c = 150).

IV. MODEL VALIDATION

What is left is the validation of our analytical model. We validate it by means of simulation and care mainly about the late (stationary) phase and situations where CBP ≤ 0.1, because we assume that if CBP > 0.1 this means that QoE is too low anyway and, therefore, model accuracy is not really important for those cases.

We validate the model by means of two series of experiments and observed good agreement between the analytical model and the simulation results. Therefore, we consider the analytical model as being sufficiently realistic.
Of course, our validation phase is limited by the fact that we do not have any access to measurements regarding IPTV service availability in vehicular networks because those systems currently do not yet exist. So, we find it acceptable to rely on IPTV service availability predictions based on a detailed and (hopefully) sufficiently realistic simulation model.

A. Series I of Validation Experiments

In Series I, we changed N_c (the number of users in the cell) and kept N (the number of channels available) and BW_c (the maximum number of channels that can be broadcasted at the same time) constant per set of experiments, with N = 50 and BW_c = 30 for *set 1* of Series I and N = 100 and BW_c = 40 for *set 2*. As can be seen in Table II, the analytical model and the simulation model are matching quite well with a few minor outliers at N_c = 200 in both sets. Also, the values of the analytical model in set 1 do not increase as fast as the values of the simulation model (with increasing N_c).

B. Series II of Validation Experiments

In Series II, we kept the number of users per cell constant (N_c = 300) and changed N (the number of channels available) and BW_c (the maximum number of channels that can be broadcasted at the same time). We, again, observe a good agreement between the analytical model and the simulation results, with a few minor outliers at higher values for N, where the analytical model is a close upper bound; for details regarding the deviations, cf. Table III.

 ISBN: 978-1-61208-398-8

TABLE II. SERIES I OF VALIDATION EXPERIMENTS

Series I: Set 1 N=50 BW_C=30				
	CBP		**Deviation**	
N_C	**AM**	**SM**	**Relative [%]**	**Absolute**
100	0,0011	0,0024	-118,18182	-0,0013
200	0,0506	0,0339	33,003953	0,0167
300	0,0577	0,0549	4,8526863	0,0028
400	0,0578	0,0615	-6,4013841	-0,0037
500	0,0578	0,0649	-12,283737	-0,0071

Series I: Set 2 N=100 BW_C=40				
	CBP		**Deviation**	
N_C	**AM**	**SM**	**Relative [%]**	**Absolute**
100	0,00008	0,00009	-12,5	-0,00001
200	0,068	0,0327	51,911765	0,0353
300	0,0846	0,0642	24,113475	0,0204
400	0,0843	0,0832	1,3048636	0,0011
500	0,0843	0,0869	-3,084223	-0,0026

TABLE III. SERIES II OF VALIDATION EXPERIMENTS

Series II N_C=300				
	CBP		**Deviation**	
N, BW_C	**AM**	**SM**	**Relative [%]**	**Absolute**
20, 10	0,1157	0,1364	-17,891098	-0,0207
20, 15	0,0456	0,0501	-9,8684211	-0,0045
50, 20	0,1099	0,1222	-11,191993	-0,0123
50, 30	0,0577	0,0529	8,3188908	0,0048
75, 30	0,0942	0,0956	-1,4861996	-0,0014
75, 40	0,0607	0,0424	30,14827	0,0183
75, 50	0,0074	0,0057	22,972973	0,0017
100, 50	0,0473	0,0251	46,934461	0,0222
100, 60	0,0016	0,0017	-6,25	-1E-04
150, 50	0,0889	0,0533	40,044994	0,0356
150, 60	0,0381	0,0182	52,230971	0,0199
150, 70	0,0011	0,0009	18,181818	0,0002

V. A GENERALIZED APPROACH TO PREDICT CHANNEL BLOCKING PROBABILITIES

In the following, our goal will be to use our analytical model, presented in Section III, to predict with only very little expenditure the availability of IPTV services in VANETs. In particular, our approach should cover a broad spectrum of traffic situations and of network technologies used to establish the access network for vehicle to RBU communication and, last not least, it should also cover numerous characteristics of the IPTV service offered.

Calculation of CBP based on our analytical model yields to the following formula:

$$CBP = P^* \bullet \sum_{i>BW_c}^{N} p_i ,$$

and this shows that CBP can be seen as a product of only two terms T_1 and T_2 with

$$T_1 \triangleq P^* \quad \text{and} \quad T_2 \triangleq \sum_{i>BW_c}^{N} p_i$$

If we fix the value of the parameter θ in the Zipf distribution used to model IPTV user behavior, it becomes evident that

$$T_1 = T_1(N, N_c, BW_c) \quad \textit{and} \quad T_2 = T_2(N, BW_c).$$

Therefore, it is possible to characterize T_1, as well as T_2 by means of elementary sets of curves. Moreover, T_1 is a general upper bound for CBP because

$$T_1 = P^* > P^* \bullet \sum_{i>BW_c}^{N} p_i = CBP$$

This is why the sets of curves related to term T_1 (resp. P*) are of particularly strong interest. Similarly, T_2 is an upper bound of CBP, too.

A. Characterization of term T_1, i.e., P*

Here, we want to investigate the influence of the

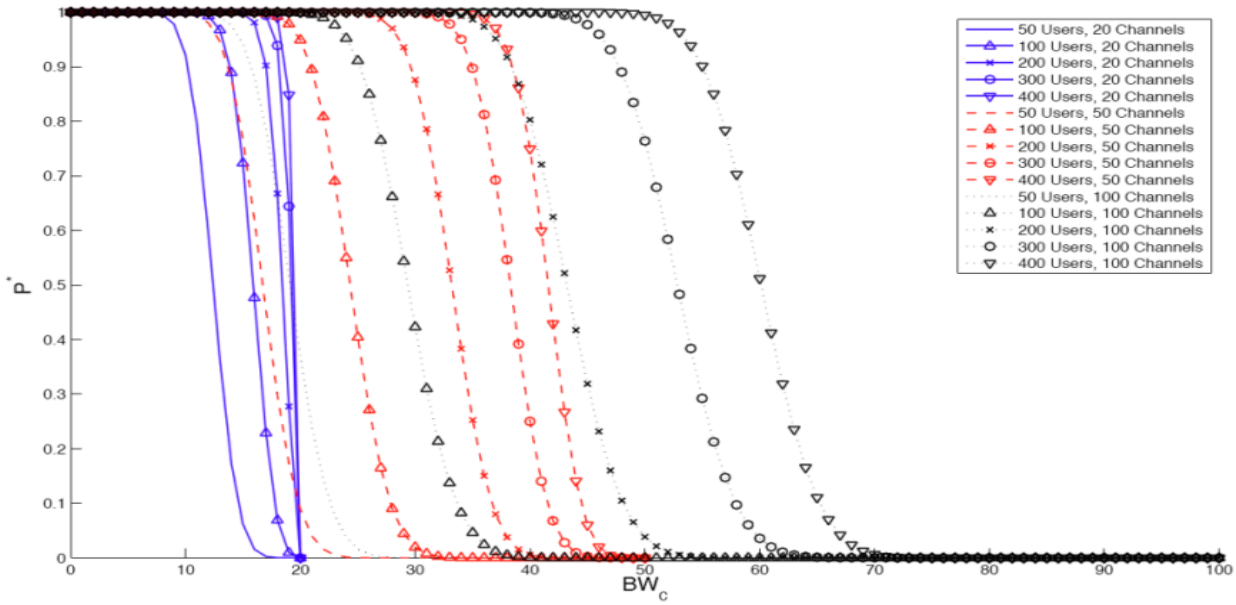
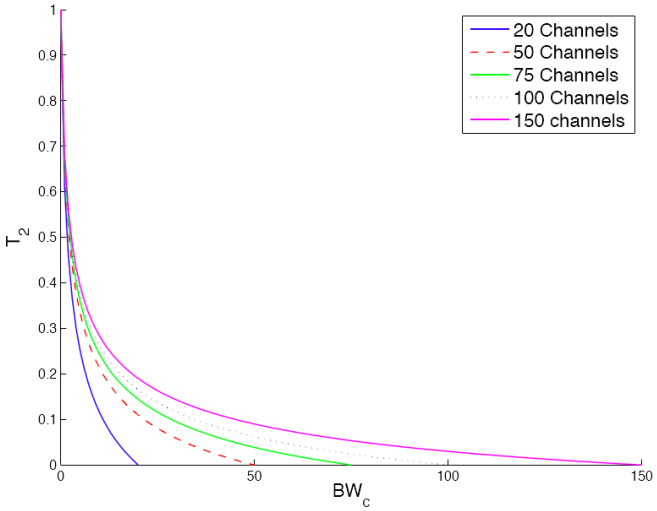


Figure 3. P* as a function of BW_c for different values of N and different cell populations N_c of IPTV users



F igure 4. T_2 as a function of BW_c for different values of N

available bandwidth BW_c on P* assuming that a certain number N of channels is offered and that the number N_c of IPTV users in the cell varies. In this study of P*, we assume $N \in \{20, 50, 100\}$ because $N = 20$ presents a small, $N = 50$ a medium and $N = 100$ a quite large number of channels offered.

Moreover, we suppose $N_c \in \{50, 100, 200, 300, 400\}$ because in realistic scenarios (e.g., for $\alpha = 0.05$) one nearly always will have no more than 400 IPTV users in a single cell (cf. below). Evidently, variation of BW_c only makes sense in the interval $[1, N]$.

Fig. 3, e.g., directly shows that if $N = 100$ channels are offered, spending a bandwidth $BW_c = 70$ for IPTV will lead to a negligible value of P* and, therefore, also to a negligibly small CBP for all realistic cell populations considered by us ($N_c \leq 400$). And even a bandwidth $BW_c = 65$ reserved for IPTV will ensure that CBP < 10 % holds, if again $N_c \leq 400$ can be assumed.

B. Characterization of term T_2

As T_2 is no longer dependent on N_c, investigations concerning this term become even more straight-forward than for T_1. In particular, the dependency of T_2 on the bandwidth BW_c reserved for IPTV can be directly depicted for a given value of N.

TABLE IV. N_c AS A FUNCTION OF $\bar{d}$ AND C_r

$\bar{d}$ \ C_r	1 km	3 km	5 km	10 km
5 m	120	360	600	1200
10 m	60	180	300	600
20 m	30	90	150	300
50 m	12	36	60	120
100 m	6	18	30	60

Figure 4 shows those dependencies for $N \in \{20, 50, 75, 100, 150\}$. This figure provides in-depth insight regarding the difficult decision of how much bandwidth should be spent for a given number N of offered channels. As examples, let us look at the case of $N = 20$ where it seems to be a good idea to choose $BW_c \geq 18$ (at least), for $N = 75$ a bandwidth of at least $BW_c = 50$ seems to be desirable and for $N = 150$ a chosen bandwidth of $BW_c \leq 80$ seems to be quite risky.

C. Expected number of IPTV users in a cell

The number N_c of IPTV users to be expected in a cell will just depend on:

- average distance $\bar{d}$ between two adjacent vehicles (driving in the same lane), where the avg. is taken over all lanes
- the n^{o} of lanes per direction (k)
- the probability that in a vehicle IPTV is used (α)
- the cell radius (C_r).

In particular, N_c can be easily determined as follows:

$$N_c = \alpha \bullet 2k \bullet 2C_r / \bar{d}$$

If we set $\alpha = 0.05$ and $k = 3$ to be constant and if we vary $\bar{d} \in \{5\text{m}, 10\text{m}, 20\text{ m}, 50\text{ m}, 100\text{ m}\}$ and assume cell radiuses of $C_r \in \{1\text{ km}, 3\text{ km}, 5\text{ km}, 10\text{ km}\}$, we get N_c values as depicted by Table IV. We see that with our assumptions, which we consider to be quite realistic, the value of N_c varies between 6 and 1200. We also can observe that rather

TABLE V. CHANNEL BLOCKING PROBABILITY (*CBP*) FOR DIFFERENT COMBINATIONS OF N, BW_c VALUES AND DIFFERENT N_c VALUES

CBP									
N_c \ (N, BW_c)	50	75	100	125	150	200	300	400	500
(20, 15)	0,0028	0,0175	0,033	0,0411	0,0442	0,0455	0,0456	0,0456	0,0456
(50, 20)	0,0098	0,0712	0,1043	0,1094	0,1099	0,1099	0,1099	0,1099	0,1099
(50, 30)	0	0,00002	0,0011	0,0086	0,0243	0,0506	0,0577	0,0578	0,0578
(75, 30)	0,000001	0,0011	0,0193	0,0585	0,0843	0,094	0,0942	0,0942	0,0942
(75, 50)	0	0	0	0	0	0,00003	0,0074	0,0302	0,038
(100, 50)	0	0	0	0	0,00001	0,0024	0,0473	0,0616	0,0619
(100, 60)	0	0	0	0	0	0	0,0016	0,0227	0,0414
(150, 60)	0	0	0	0	0	0,0002	0,0381	0,0717	0,073
(150, 70)	0	0	0	0	0	0	0,0011	0,0293	0,0563
(150, 80)	0	0	0	0	0	0	0	0,0009	0,0162

 ISBN: 978-1-61208-398-8

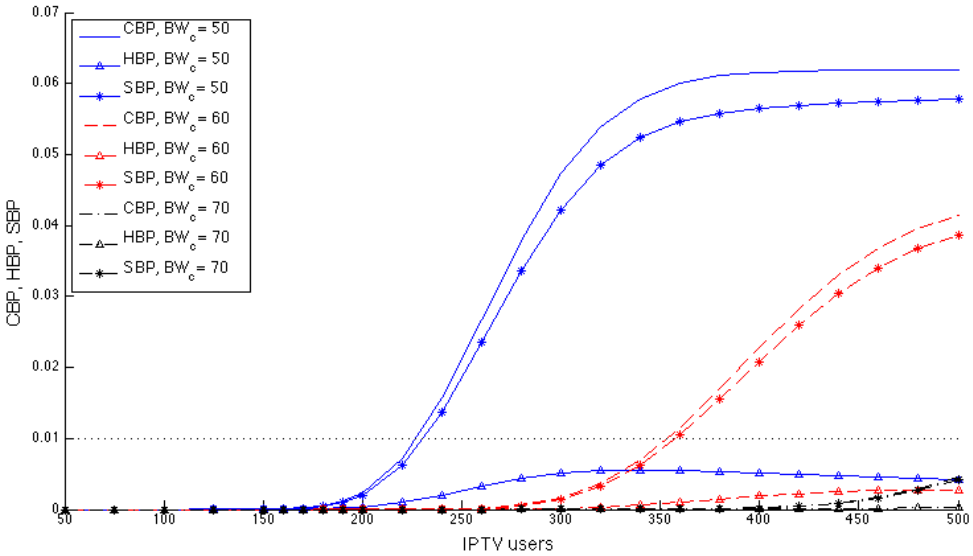


Figure 5. *CBP*, *HBP* and *SBP* for different values of BW_c in dependence of N_c

different combinations of parameter values will lead to the same value of N_c which facilitates the characterization of P* and thus also of CBP.

D. Straight-forward calculation of CBP for numerous scenarios of IPTV in VANETs

Combining the results achieved in this section up to now, we are able to propose a generalized proceeding which allows us to predict CBP for nearly any scenario of interest with nearly negligible expenditure (if we compare this with a CBP prediction based on simulation models for assessing IPTV availability in VANETs).

In particular, Table IV showed us which N_c to assume to be realistic and the results of Fig. 3 and 4 can be directly combined (i.e., T_1 and T_2 can be multiplied) to determine CBP. Table V contains CBP predictions based on our analytical model for numerous scenarios of IPTV in VANETs. The results of Table V cover a broad spectrum of traffic situations (low, medium and high traffic load up to traffic jam), of access network technologies used having an impact on C_r and BW_c and of characteristics of the IPTV service (e.g., number N of channels offered).

To summarize, the results obtained in this section can allow one to significantly improve the understanding of the main factors and their mutual dependencies which influence IPTV availability in VANETs.

VI. CASE STUDY

In the previous section, we have shown how it is possible to determine CBP just as a function of N, N_c and BW_c, where, of course, N_c itself is a function of $\overline{d}$, C_r, k and α. We now want to indicate how the handover- and the switching-induced blocking probabilities HBP and SBP can be determined based on CBP.

A. Calculation of $\#ho_{ph}$

Let $\#ho_{ph}$ denote the total number of handovers per hour of all vehicles using IPTV and leaving a given cell. We assume a mean speed of those vehicles of $\overline{v}$ and a mean distance between adjacent vehicles of $\overline{d}$, a cell radius C_r, k lanes per direction, as well as an IPTV watching probability of α. With these assumptions we can directly calculate N_c (cf. Section V.C.).

$\#ho_{ph}$ can be determined in a straight-forward manner as follows:

$$\#ho_{ph} = \alpha \bullet 2k \frac{\overline{v}\,[\frac{km}{h}]}{\overline{d} \bullet 10^{-3}\,[km]} = \alpha \bullet 2k \frac{\overline{v}}{\overline{d} \bullet 10^{-3}}\,[\frac{1}{h}]$$

B. Calculation of $\#sw_{ph}$

Let $\#sw_{ph}$ denote the total number of switching events per hour of all N_c vehicles using IPTV in a given cell. Let us assume a mean time Δt [min] between two successive channel switching events, where $\Delta t = 3$ [min].

Then, $\#sw_{ph}$ can be determined as follows:

$$\#sw_{ph} = \frac{60}{\Delta t} \bullet N_c\,[\frac{1}{h}]$$

C. Calculation of HBP and SBP

HBP can be determined based on $\#ho_{ph}$, $\#sw_{ph}$ and CBP as follows:

$$HBP = (CBP \bullet \#ho_{ph})/(\#ho_{ph} + \#sw_{ph})$$

Correspondingly:

$$SBP = (CBP \bullet \#sw_{ph})/(\#ho_{ph} + \#sw_{ph})$$

D. Case Studies

Let us now apply the formulae for HBP and SBP to concrete scenarios for VANETs offering IPTV service.

We assume a medium traffic situation with $k = 3$, $\overline{d} = 50$ m, $\overline{v} = 120$ km/h (averaged over all 2k lanes), $C_r = 5$ km, $N = 100$, $\alpha = 0.05$, $BW_c \in \{50, 60, 70\}$.

Fig. 5 shows the values of CBP, SBP and HBP in dependence of N_c for the 3 values assumed for BW_c. Of course, increasing N_c just corresponds to increasing the cell radius C_r. Curves for SBP have been included in Fig. 5 because of our trial to facilitate result interpretation, though the SBP curves are a direct consequence of the two other ones as SBP is just the difference SBP = CBP – HBP.

Among others, Fig. 5 shows that for BW_c = 70 CBP remains rather small for all values of N_c considered and even for N_c = 500 the value of CBP still remains well below 0.005. Results such as in Fig. 5 may be highly valuable for a provider of an IPTV service in VANETs because they allow one to answer questions such as:

- How many users are acceptable in a cell if a certain bandwidth BW_c is available for IPTV and we want to keep CBP below a certain threshold? The threshold for CBP could be 0.01 as it is indicated in Fig. 5 and we can observe that if BW_c = 50 then $N_c \leq 220$ is still acceptable or if BW_c = 60 then $N_c \leq 350$ will still lead to the desired QoE.
- How strongly will HBP, SBP and CBP depend on N_c (N_c here mainly being a function of the cell size)?
- How much does an increase of BW_c help in reducing the blocking probabilities?
- What proportion of CBP is due to handover respectively switching events, i.e., how much does handover-related blocking "hurt" QoE?

VII. SUMMARY AND OUTLOOK

The goal of this paper has been to evaluate the availability of IPTV services in VANETs based on an analytical modeling approach. The analytical model elaborated allows one to predict TV channel availability in a quite flexible and highly efficient manner (if, e.g., compared to simulation models). Our comprehensive validation studies indicate that the analytical model achieves a satisfactory degree of validity. The general proceeding for making use of the model demonstrates that studies are possible in a straight-forward manner covering strongly different scenarios. The proceeding proposed is primarily based on a highly relevant discovery made by us, namely, that we realized that the terms T_1 and T_2 , the product of which yield CBP, (cf. Section V) are both upper bounds of CBP. The case studies illustrate how the general proceeding for efficient experimentation can lead to a valuable understanding of the main factors influencing (switching- as well as handover-induced) TV channel blocking probabilities in VANETs.

In our future research, we plan to elaborate an algorithm which allows one to reduce the probability of handover-induced channel blockings (at the expense of an increasing number of switching-induced blockings). Such an algorithm seems to be highly desirable in order to alleviate the strongly negative impact, which handover-induced blockings may have onto QoE in the provisioning of IPTV services in VANETs.

REFERENCES

[1] A. Abdollahpouri, QoS Aware Live IPTV Streaming Over Wireless Multi-hop Networks. Shaker Verlag 2012.

[2] A. Abdollahpouri, B. E. Wolfinger, J. Lai, and C. Vinti, Modeling the Behavior of IPTV Users with Application to Call Blocking Probability Analysis. Praxis der Informationsverarbeitung und Kommunikation 2012, 35 (2), pp. 75-81.

[3] A. Baiocchi and F. Cuomo, Infotainment services based on push-mode dissemination in an integrated VANET and 3G architecture. Journal of Communications and Networks, April 2013, 15 (2), pp. 179-190.

[4] W. L. Dunn and J. K. Shultis, Exploring Monte Carlo Methods. Elsevier Science 2011.

[5] FOCUS Online: 80 Prozent der Neuwagen mit Internet. www.focus.de/auto/news/revolution-im-auto-bis-2016-haben-80-prozent-der-neuwagen-internet_id_4216763.html (last access: Dec. 18, 2014).

[6] M. Krohn, R. Daher, M. Arndt, and D. Tavangarian, Aspects of roadside backbone networks. 1st Internat. Conf. Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, Wireless VITAE 2009, pp.788-792.

[7] J. Lai, Evaluation and Improvement of TV Channel Availability for IPTV Services. Shaker Verlag 2012.

[8] S. Malkos, E. Ucar, and R. Akdeniz, Analysis of QoE key factors in IPTV systems: Channel switching. 5th Internat. Conf. on Application of Information and Communication Technologies (AICT2011), October 2011, Baku, Azerbaijan, pp. 1-5.

[9] S. Möller and A. Raake, (eds.), Quality of Experience: Advanced Concepts, Applications and Methods. Springer 2014, pp. 11-35.

[10] S. Momeni, J. Lai, and B. E. Wolfinger, Availability Evaluation of IPTV Services in Roadside Backbone Networks with Vehicle-to-Infrastructure Communication. 9th Internat. Wireless & Mobile Computing Conference. IWCMC 2013, IEEE Cagliari, Sardinia, Italy, 2013, pp.1727-1732.

[11] S. Momeni and B. E. Wolfinger, Availability Evaluation of IPTV in VANETs with different types of access networks. EURASIP Journal on Wireless Communications and Networking, Springer Open Journal, 2014: 117 (15 July 2014).

[12] M. E. J. Newman, Power Laws, Pareto Distributions and Zipf's Law. Contemporary Physics 2005, 46 (5), pp.1-3.

[13] RFC 1112: Host Extensions for IP Multicasting. August 1989.

[14] R. Y. Rubinstein and D. P. Kroese, Simulation and the Monte Carlo Method. 2nd ed., Wiley 2007.

[15] J. P. Urrea Duque and N. Gaviria Gomez, Quality assessment for video streaming P2P application over wireless mesh network. XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA2012), September 2012, Antioquia, Colombia, pp. 99-103.

[16] B. E. Wolfinger, A. Hübner, and S. Momeni, A Validated Analytical Model for Availability Prediction of IPTV Services in VANETs. Electronics 2014, 3, pp. 689-711; doi: 10.3390/electronics3040689.

 ISBN: 978-1-61208-398-8

Prediction Metrics for QoE/QoS in Wireless Video Networks for Indoor Environmental Planning: A Bayesian Approach

André Augusto Pacheco de Carvalho
Laboratory of Computation and Telecommunications
Federal University of Pará, UFPA,
Belém, Brasil
e-mail: andrepcarvalho@gmail.com

Simone da Graça de Castro Fraiha
Laboratory of Computation and Telecommunications,
Institute of Exact and Natural Sciences, Federal University of Pará Belém, Brasil
e-mail: fraiha@ufpa.br

João Victor Costa Carmona
Laboratory of Computation and Telecommunications
Technology Institute, Federal University of Pará, UFPA,
Belém, Brasil
e-mail: victorjvcc@gmail.com

Hermínio Simões Gomes
Laboratory of Computation and Telecommunications,
Institute of Exact and Natural Sciences, Federal University of Pará Belém, Brasil
e-mail: herminio@ufpa.br

Jasmine Priscyla Leite de Araujo
Laboratory of Computation and Telecommunications,
Technology Institute, Federal University of Pará
Belém, Brasil, Bolsista CNPq-Brasil
e-mail: jasmine.araujo@gmail.com

Gervásio Protásio dos Santos Cavalcante
Computer and Telecommunication Engineer Faculty
Federal University of Pará, UFPA,
Belém, Brasil
e-mail: gervasio@ufpa.br

***Abstract*—The evolution of applications on wireless networks has grown in recent years due to the growth in the number of users of mobile phones, tablets, and other. The availability of demanding services, such as video transmission affect the Quality of Experience (QoE) and Quality of service (QoS) provided domestic and commercial users, it has stimulated the study of new techniques of management of network resources, always having as objective to provide high quality services to an increasingly demanding customer. This paper presents a methodology of Artificial Intelligence, using the technique of Bayesian networks, as a hybrid evaluation strategy by analyzing the behavior of QoE and QoS metrics, in designing wireless LAN. So, by manipulating the basis of Bayesian networks, it was possible to find satisfactory results for the proposed solution helps the planning of wireless networks in indoor environments, which does not exclude the possibility of using this approach for other situations.**

Keywords-QoE/QoS; measurements; simulation; Bayesian networks; wireless networks.

I. INTRODUCTION

Currently, there is an increasing need for bandwidth services due to the demand for greater high-speed mobility and available services, any time, any place, anywhere. It should be noted that there is a widespread access to wireless networks and the use of multimedia applications such as Voice over IP (VoIP), and online video games (data and video).

The International Telecommunication Union (ITU) report [1] indicates that this increase includes network bandwidth usage. This has involved observation and improved planning in Wireless Local Area Networks (WLAN), so that bottlenecks and/or overloading can be avoided.

Studies based on measurements can provide more precise results than studies based on simulations or modeling [2].

In [3], there is a visualization tool for wireless network parameters; in this case, physical layers for the network, where the data are also collected through measurements in an indoor environment. In [4], a study was carried out on wireless and digital television transmissions in networks. Metric data of the physical layer and application were gathered with the aim of creating a cross-layer approach to model quality loss through empirical equations.

In this study, a combination of measuring and simulation (through the use of Matlab® software [5]) is proposed. This is carried out by means of modeling with the aid of Bayesian networks to represent wireless network and predict the behavior of Quality of Experience or services offered to the user that differ from [3], where the QoS was only evaluated for a VoIP application.

An evaluation will be conducted of several Quality of Experience (QoE) and Quality of Service (QoS) metrics [6], some taken from the measuring process and others originating from the data handling.

The study is structured in the following way: Section II presents the related work. In Section III, the environment where the measurements were carried out is examined. Section IV describes the methodology employed and Section V shows the results obtained. Section VI discusses the conclusions of the study.

 ISBN: 978-1-61208-398-8

II. RELATED WORK

This work has as a differential of other article published, consideration of the parameters of QoS and QoE to assist in planning design of wireless networks for indoor environments.

As we can see from Wu et al. [7], the beginning of the form study in this area of planning mainly uses techniques of neural networks and genetic algorithm to perform the forecast/prediction with some QoS/QoE metrics, never both.

Dimitriou et al. [8] makes use of physical layer parameters such as average power, signal error and signal, interference noise (SINR).

Fraiha [9] proposed a methodology for projects in wireless local area networks optimized by using a model of loses of signal power and measures considered QoS parameters, always aiming to maximize coverage with the least number of Access Point (PA) to be installed in a given indoor environment by using a technique of optimization of multi-objective genetic algorithms, proposing a propagation model that allows simulating the likelihood of receiving the signal in the environment studied.

This paper presents a new approach in designing indoor networks; prediction considering the QoE and QoS parameters to assist in the planning of wireless networks for indoor environment.

III. THE MEASUREMENT SETTING

The classroom building of the Federal University of Pará (UFPA) was used as the setting for carrying out the measurement campaign, as illustrated in Figure 1. It is built of bricks and concrete and has glass windows on the side and a corridor. The chairs and tables inside the rooms are made of plastic and metal.

The measurements were conducted on the second floor. The size of the building is 40x11 meters. This floor has six rooms, each measuring 6x8 meters.

Figure 2 shows a diagram of the classrooms where the measurements took place. The transmitter location is shows together with the points that were analyzed (in red).

Figure 1. The classroom building: classrooms and side corridor

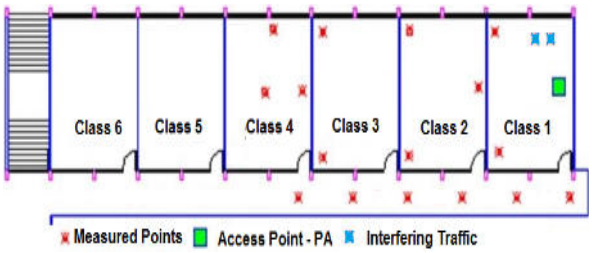


Figure 2. Model of the setting and location of the points analyzed (red) and transmitter location (PA).

IV. METHODOLOGY

The methodology employed for this study followed a number of stages: collection of the results of the experiment - these results are used to gain entry to the Bayesian networks [3], so that probability maps can be generated. After this, these maps are converted into information to allow an analysis of the behavior of the QoE/QoS metrics in wireless networks and to assist in the planning of the indoor networks. Figure 3 shows a flowchart of the methodology employed in this study.

In planning wireless networks there is a mathematical formula that proves of installation exact PA, so in the methodology of this article was carried out measurements for the construction of measures to generate Bayesian network, to conduct training and validation.

Several tools were used in order to control the traffic characterization, mapping the wireless network created, and can perform the analysis while the transmitter sent the video to the recipient.

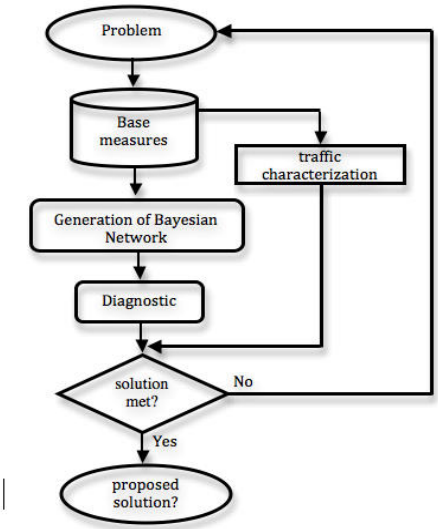


Figure 3. Flowchart of the Methodology

A. Materials Used

The components of the interconnected network were as follows: four notebooks, two being used for the video transmission and reception respectively, and the remaining

 ISBN: 978-1-61208-398-8

two to generate competing traffic in the network by using the Windows 7 operating system.

The configurations of the computers and PA that were employed, were as follows:

- Two computers with a processor: core I5 2.4 GHz; 4 GB RAM memory, HD of 500 GB; on-board video card with up to 512 MB of shared RAM memory and Windows 7 64 bit operating system;
- Two computers with AMD A6 3410mx processor; video radeon HD graphics 1.6 GHz card; 4 GB RAM memory; HD of 500 GB; and Windows 7 64 bit operating system;
- An Evo-W301AR (SIROCO) Router Model, Channel 3 (2422 MHz) - mode: 802.11g – channel width 20 MHz, maximum transmission rate 54 Mbps and Transmission power 20 dBm.

Figure 4 below shows the layout of the computers in class 1, where some of the devices used were installed.

B. *Measurements*

A WLAN network was assembled to evaluate the communication and video transmission. Evalvid [10] software was employed for the video transmission and reception; this consists of a set of tools that allow the network to be evaluated by obtaining performance metrics such as jitter, end-to-end delay, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM).

These were obtained by making a comparison between the video frames received and the video reference frames. Figure 4 shows the two frames - the frames degraded by the effects of transmission (left part) and the original frame without degradation, (right part).

In the case of the SSIM metrics, the index varies from 0 to 1, where the closest to 1, means a greater quality and indices closest to 0, means a lowering of quality. By analogy, the PSNR, is in a scale that ranges from 0 to 50, where the higher the value, the greater the inability of the user to detect failures in the video.

Some metrics like PSNR and SSIM are calculated through a set of frames. Other metrics like jitter and end-to-end delay are calculated through the time needed for the consecutive frames to reach the receptor. As well as these metrics, at each reception point, the Receiver Signal Strength Indicator (RSSI) was measured by means of the WirelessMon tool [11].

Traffic simulation (T-Iperf and R-Iperf) constant [12] was generated in the network with the aim of creating competition for bandwidth between the traffic caused by Iperf and the transmission of video traffic. This made it possible to characterize the network that was the nearest possible to a real network where competitive services are made available by the band.

The videos were transmitted in the MPEG-4 Widescreen 16:9 format with a resolution of 1920x1080p so that they complied with the digital TV coding Standards.

During the measurement procedure, a methodology was employed to ensure there was no divergence in the way the data were collected. It can be described as follows: the PA was installed in the first room together with both the computers which carried out the video transmission and the computers that generated the network traffic constant with the T-Iperf e R-Iperf measurement tools. Only the video receptor that remained moved around between the points and this meant that the video could be received and reconstructed and it was thus possible to evaluate the quality of the transmission.

Figure 4. Example of frame reception failures (above) and the original video frame (below).

The sending of the video and the measurement of the signal strength generated by the PA was carried out for each point shown in Figure 2. After the dispatch, a receptor file was made which enabled the video to be reconstructed by checking the quality of the data transmitted to determine if there had been any packet losses during the transmission.

C. *Computational Modeling and Simulation (Bayesian Networks)*

At the end of the measuring, several files were compiled and on the basis of this data, the modeling was carried out to enable us to create a Bayesian network.

Bayesian networks are probabilistic graphical models for the representation of knowledge in areas where there is uncertainty. These models can be represented in two way:

- **Qualitative**: representing the dependencies between the nodes;
- **Quantitative:** represented by conditional probability tables.

As a result, the evaluation can be carried out in terms of the probability of these dependencies [13][14]. These components form an efficient representation of the joint probability distribution of the X variables in a given domain [15].

Since a network is generated, it is possible to predict its future behavior through the choice of an interval for a given attribute. When this interval is selected, the values are propagated to the nodes, which is called the Bayesian inference [14].

After the measurements, numerous files were generated; these files contain data that generated the basis of measures. However, due to signal coverage, we could not obtain all the necessary measures for the entry of Bayesian network.

 ISBN: 978-1-61208-398-8

However, with the use of Matlab®, it has been possible to perform the database extension measures. Dealing with a real existing base, we could create a Bayesian network.

Despite having generated multiple files, scientifically measured points were to be based on an analysis of a methodology. However, due to coverage area of PA and the size of the building, we could not do more measurements; so, as a solution to extend the base, Matlab® has been used.

A program was created that uses the feature of Matlab® of a micro type artificial neural network Radial Basis Function (RBF), the function *newgrrnn*, which implements the generalized recurrence, where a Gaussian function [2] with crests near the points measured is simulated. So, the points obtained from regression [9] has underestimated values and values well estimated to close points measured, making the gradual expansion of the points so that there was a sufficient database for Bayesian network was applied.

Furthermore, in the program, there was the need of a softening in the generation of items, by using the parameter in the Matlab ® called *spread*, for which the creation of new point was approximately, so the regression cited would point creation by simulating the real scenario, where points close to the measured point stay good and far from the worst gets worse.

Figure 5 shows the network generated with the QoS and QoE metrics and the metrics for physical layer, distance and RSSI. Figure 6 shows the inference of best distance. The probability values of the other nodes are altered, which confirms the propagation.

This was exemplified by analyzing the RSSI metric when it was observed that there is a 81% probability of being between -60.0 and -49.0 dBm when the inference for the distance metric is at its best value, with points close to the access point of 5.36 to 6.06 meters. The PSNR value for the same inference has 65% of probability of being between 24.55 and 36.24dB. The delay has almost 100% of probability of being between 151 and 591 ms. Jitter has a 99.0% probability of being between entre 0.08 and 1.83s. Finally, the SSIM has a probability of 65% of being between 0.75 and 0.868.

Figure 7 gives an example of another inference that can be made with regard to the worst distance. When the RSSI metric is analyzed, it can be seen that it has an 86% probability of being between -64.0 and -79.8 when the inference of the metric distance is at its worst value, if the worst value is considered to be the points most distant from the access point of 12.55 to 24.13 meters. The PSNR value for the same inference has a 62.1% probability of being between 13.33 dB to 19.89 dB.

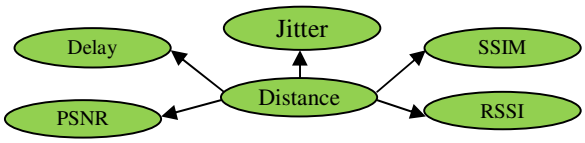


Figure 5. Bayesian Network

The Delay has more than 50% of probability of being greater than 194 ms. Jitter has a 61.3% probability of being between 1.834 and 3.69 seconds. Finally, the SSIM has a probability of 62.1% of being below 0.65. Figure 8 and Figure 9 can be analyzed for a visual inspection of the inferences carried out.

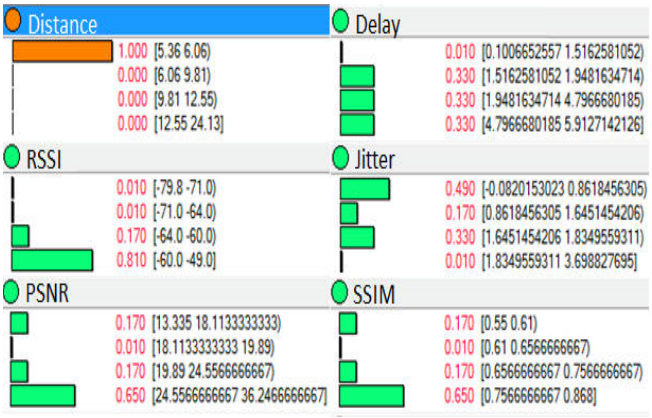


Figure 6. Inference of best distance

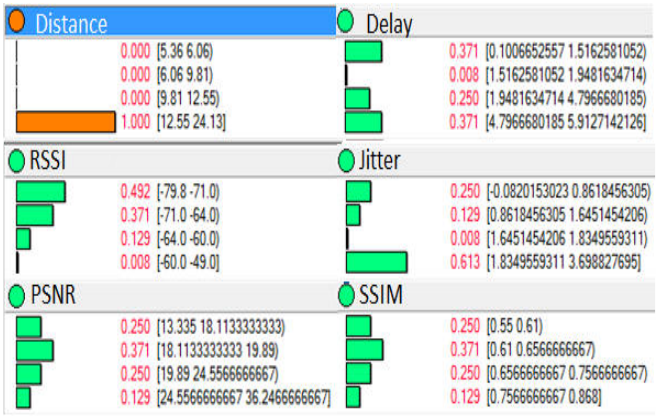


Figure 7. Inference of worst distance

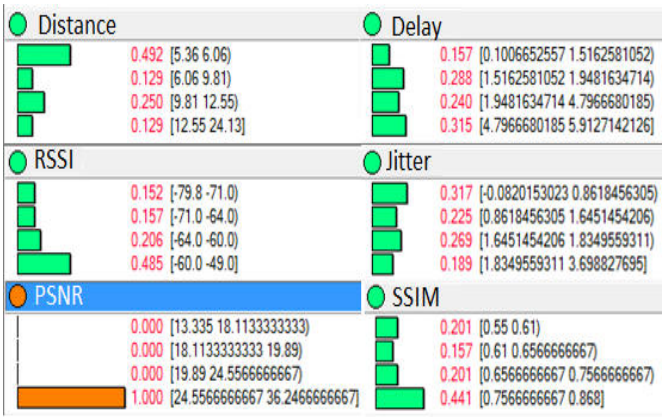


Figure 8. Inference of best PSNR

 ISBN: 978-1-61208-398-8

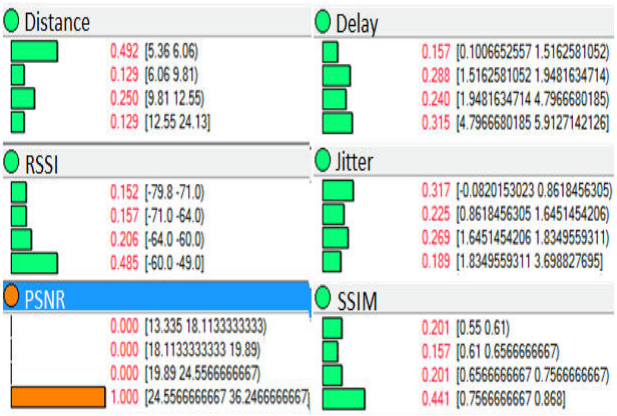


Figure 9. Inference of worst PSNR

With the aid of the conditional probability tables obtained from the Bayesian network and variance tables, a simulation was conducted on the basis of this mathematical modeling and adapted to QoE metrics by means of Matlab. Finally, the values of the distance metrics can be visualized to analyze/predict the behavior of the video streaming network.

V. RESULTS

The heat maps presented follow the same reasoning of the measured building plan; for this reason, from Figure 10 to Figure 14, the X axis represents the distance in meters and the simulation also represents the layout of the ¨pavilion¨ (inner part of the building) described in Section II. The Y axis represents the width of the classroom and the colors represent the degree of probability of the metric. The reference-point can be considered to be at the beginning of the X axis at a distance of 40 meters, that is it passes through the pavilion in an opposite direction to that shows in the diagrams. All the distance values shows in this section follow this reference-point.

With regard to the SSIM metric, the greater it is, the better according to Figure 10, and a Bayesian simulation with measurement data shows a behavior above 0.7, even up to 10 meters from class 1.

In the case of the delay metric, Figure 11 shows a reasonable value that corresponds to less than 250 ms at a distance of approximately 10 meters too. In Figure 12, jitter follows the previous metrics and shows reasonable values in up to 10 meters with values less than 0.8 second.

PSNR is most often used to measure the quality of reconstruction of loss compression codec (example: for image compression). The sign, in this case, are the original data, and the noise is the error introduced by compression. When comparing compression codec, PSNR is an approach to the human perception of quality reconstruction.

In Figure 13, PSNR also has satisfactory values up to 10 meters with values that are approximately 28 or more. In Figure 14, the RSSI also has values higher than 10 meters, above 60 dB.

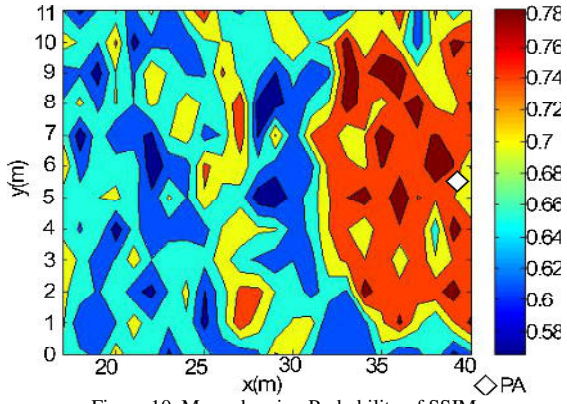


Figure 10. Maps showing Probability of SSIM

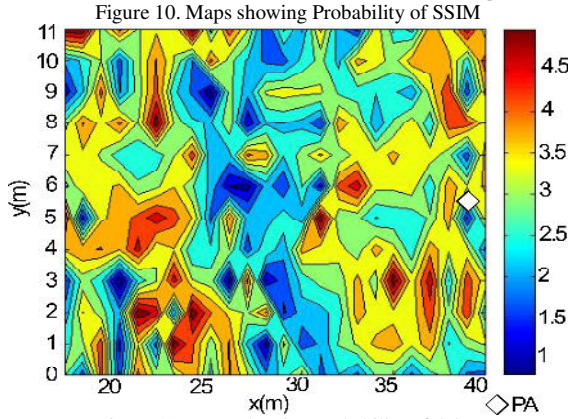


Figure 11. Maps showing probability of delay

As in visual terms, the metrics had a suitable performance at a distance of up to 10 meters from the Access point; in this case, it is suggested that in the planning stage, other PAs are installed from this point, so that the other classrooms can be provided with a satisfactory number of applications that use video.

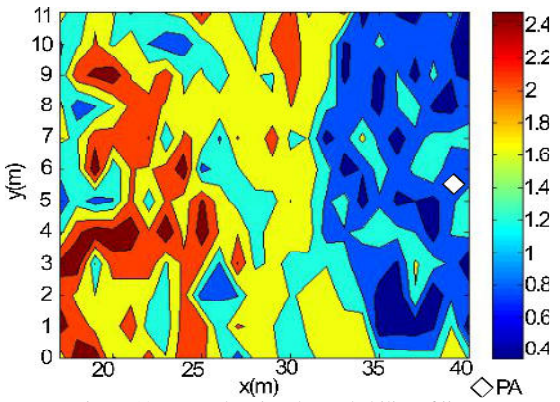


Figure 12. Maps showing the probability of jitter

This decision-making is based on the use of video to meet the needs of metrics that measure performance within a specified standard. Thus, the behavior of metrics can be predicted and access points installed, to make improvements when complying with the requirements of the QoS/QoE parameters.

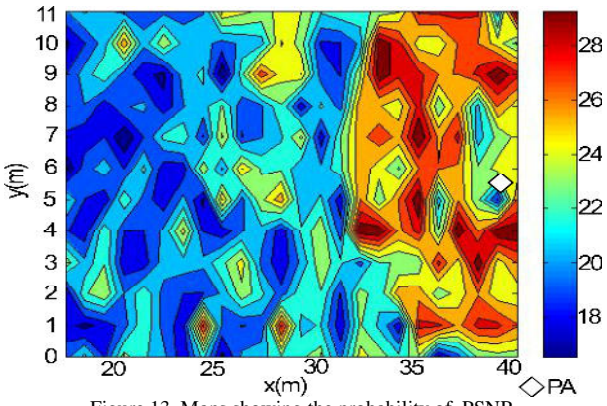


Figure 13. Maps showing the probability of PSNR

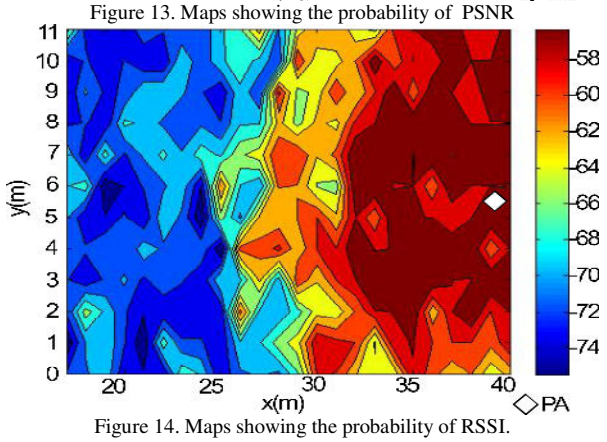


Figure 14. Maps showing the probability of RSSI.

Finally, these probabilistic maps show the use of the hybrid technique: measurements and the Bayesian approach. They can be used to analyze the behavior of wireless networks and assist in the analysis of the performance of these networks by taking account of both QoE/QoS bandwidth applications in the growth of indoor wireless networks.

VI. CONCLUSION AND FUTURE WORK

This article has outlined an empirical approach and simulation for assisting the planning of wireless networks based on video measurements in the inside environment of the facilities of UFPA.

The results suggest that the use of the proposed tool is feasible. The measured data allow the Bayesian network to make estimates in a consistent and reliable way for environments similar to those measured in this study.

In future studies, it is intended to perform more measurements campaigns, however these measurements shall be directed in other buildings, with totally different internal architecture of the architectures present in the building chosen for this dissertation.

Perform simulations with other types of network traffic, such as audio. Expand the methodology of this study for the outdoor environment, and tests with other emerging technologies, such as the 4G and 5G.

At the time of the confrontation of environments, would be inevitable, as a consequence the achievement also approach in the frequency range that used, as used in the frequencies of 2.4 GHz to 5 GHz, based on the IEEE 802.11 wireless networks while the 4G is used in the 700MHz frequency in Brazil.

ACKNOWLEDGMENT

This work was done with the support of CNPq - National Council for Scientific and Technological Development – Brazil. The authors also would like to thank to UFPA, and the National Institute of Science and Technology – Wireless Communications (INCT-CSF) and their research staff for their support of this study.

REFERENCES

[1] ITU-T, “Measuring the Information Society Report 2014” in http://www.itu.int/en/ITU-D/Statistics/Pages/publications/mis2014.aspx [retrieved: 03/2015].

[2] R. Jain. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling, Publisher Wiley, 1991, pp. 418-425.

[3] J. Araujo, J. Rodrigues, S. Fraiha, H. Gomes, N. L. Vijaykumar, G. Cavalcante, and C. Frances, “A Visualization Tool for Analyzing the Design of Physical and Network Layers Parameters of a Wireless Network”, Proceedings of the Second International Conference on Advances in Computing and Information Technology, vol. 3, July 2012, pp. 291-305.

[4] B. S. L. Castro, “CrossLayer Modeling of Quality of Experience for Video Transmission in OFDM Wireless Systems”, Doctoral Thesis, UFPA,Feb. 2014, pp. 20-24 in http://repositorio.ufpa.br/jspui/handle/2011/5593 [retrieved: 03/2015].

[5] J. B. Klaue, B. Rathke, and A. Wolisz, “EvalVid - A Framework for Video Transmission and Quality Evaluation”, 13th International Conference on Modeling Techniques and Tools for Computer Evaluation, Urbana, Illinois, USA, Sept. 2003, pp. 255-272.

[6] H. Chaari, K. Mnif, and L. Kamoun, “An Overview of Quality Assessment Methods of Video Transmission over Wireless Networks”, Electrotechnical Conference (MELECON), 16th IEEE Mediterranean, vol. 1, March 2012, pp. 741-744.

[7] R. Wu, Y. Lee, and S. Chen, “Planning System for Indoor Wireless Network”, Consumer Electronics, IEEE Transactions on, IEEE Consumer Electronics Society, vol. 47, Feb. 2002, pp. 73-79.

[8] A. G. Dimitriou, S. Siachalou, A. Bletsas, and J. N. Sahalos, “An efficient propagation model for automatic planning of indoor wireless networks”. Antennas and Propagation (EuCAP), 2010 Proceedings of the Fourth European Conference on. Barcelona, April 2010, pp. 1-5.

[9] S. G. C. Fraiha, “Location of Access Points in Indoor Environments in Wireless System Projects”, Doctoral Thesis, UFPA, May 2009, pp. 6-104.

[10] Matlab, Matlab version 7.0. In Technical Documentation, 2012.

[11] WirelessMon, PassMark Software, 2010.

[12] Y. Byun, S. Narayanan, S. Mott, K. Biba, J. Schwenkler, R. Osborn, and M. Morris, “Wireless Broaband Measurement in California”. 10th International Conference on Information Technology: New Generations, vol. 1, April 2013, pp. 505-509.

[13] Z. Chen, Data Mining and Uncerta in Reasoning: An Integrated Approach,Wiley-Interscience, Aug. 2001, pp. 285-289.

[14] K. B. Korb and A. E. Nicholson, Bayesian Artificial Intelligence, Crc Press, 2004, pp. 74-89.

[15] J. Pearl, Probabilistic Reasoning in Intelligent System, Morgan Kaufmann Publishers, Sep. 1988, pp. 223-232.

 ISBN: 978-1-61208-398-8

Non-Invasive Estimation of Cloud Applications Performance via Hypervisor's Operating Systems Counters

Fábio Diniz Rossi, Israel Campos de Oliveira,
César A. F. De Rose
Faculty of Informatics
Pontifical Catholic University of Rio Grande do Sul
Porto Alegre/RS, Brazil
{*fabio.diniz, israel.oliveira*}*@acad.pucrs.br, cesar.derose@pucrs.br*

Rodrigo Neves Calheiros, Rajkumar Buyya
The Cloud Computing and Distributed Systems Lab
Department of Computing and Information Systems
The University of Melbourne, Australia
{*rnc, rbuyya*}*@unimelb.edu.au*

***Abstract*—The adoption of cloud computing environments as the infrastructure of choice for computing services is growing rapidly, due to features such as scalability and pay-per-use. As a result, more pressure is put on cloud providers, which manage the underlying computing platform, to maintain the Quality of Experience of application users within acceptable levels. However, the mapping of high-level application metrics, such as response time, to low-level infrastructure metrics, such as utilization rate of resources, is a non-trivial task. Many works present monitoring of processor, memory, and network utilization. Nevertheless, the monitoring of these resources can be intrusive to the system that provides the service. This paper presents a non-invasive approach for estimating the response time of cloud applications through the mapping of Quality of Service metrics to operating system counters at the hypervisor level. We developed a model that estimates the response time of real-time applications based on Linux Operating Systems counters that presented an accuracy of 94% in our evaluation.**

Keywords-Cloud Computing; IaaS; QoE; Response time; SLA.

I. INTRODUCTION

Cloud computing and cloud storage have become the preferred methods for distributing information and online functionality over the Internet [1]. While some cloud service providers focus on providing customers with a broad range of features and services, including online shopping, search, social network, entertainment consumption, and protecting important documents, other cloud service providers focus on providing services for small businesses, large corporations, governments, and other institutions.

Most of these environments need to improve Service Level Objectives (SLOs) and meet Service Level Agreements (SLAs) in terms of availability, performance, security, and data protection [2]. This is important, as it impacts directly on the Quality of Experience (QoE) of users, who may or may not remain loyal to the offered services [3] [4].

With the aim of offering services that comply with these high level quality metrics established without excessive operational costs, cloud service providers use rapid elasticity provided by cloud computing [5]. Thus, it is possible to dynamically increase or decrease instances of virtual machines and/or compute nodes, as well as the applied quota of CPU, memory, and network bandwidth on a cloud service. Besides the obvious benefits of cost and performance for users, cloud providers can also benefit from a more efficient use of resources.

Elasticity, the capacity to dynamically change the amount of resources dedicated to a service, for more or less, is controlled via pre-defined SLAs. When these SLAs limits are exceeded, new resources are added so that the load returns to an acceptable level. When resources are underutilized, resources can be freed in order to reduce operational costs. Nevertheless, the decision on the amount of resources required to meet high-level metrics defined as SLAs is non-trivial [6], because some high-level metrics can not be easily monitored by the infrastructure. As the SLA [7] involves the definition of minimum acceptable levels of service that are expected by the customer, it is common the use of indicators for the quantitative measurement of the Quality of Service (QoS) received. Some commonly used indicators are availability, response time, and mean time between failure, among others.

As services offered by cloud service providers are accessed over the Internet, it is natural that network QoS metrics are the most important for user experience [8]. Thus, this work focuses on the response time of cloud applications. This is a metric that can influence the decision on the need for more or less resources to maintain an acceptable level of service, and it is measured by the time taken from the client request is received by the service provider until the response by the service provider is sent. However, this can not be monitored by the Infrastructure-as-a-Service (IaaS) without the risk of interference on the communication channel between the client application and the service provider. Furthermore, there are privacy issues that should be taken into account, as most users would not agree with monitoring systems running within their

 ISBN: 978-1-61208-398-8

virtual machines. In addition, there are applications that using non-standard software stacks that do not allow reliable monitoring of the response time.

Therefore, the problem that this paper addresses is how to estimate the response time of cloud applications, based only on information that can be accessed through the infrastructure, and without being intrusive in the communication channel of client applications. Accordingly, the hypothesis tested by this work is that internal operating system counters at the hypervisor level allow estimation of the client application response time based on the history of the load on the physical machine on which the application is allocated.

The aim of this work is not prediction, i.e., the objective of the proposed method is not to infer the response time before it occurs. Rather, our analysis takes place after the event occurred, and it aims at, based on the observed value, to estimate the application performance. This enables perform control actions on scalability, reacting to fluctuations of this metric.

To this end, tests to measure the response time were performed with a real three-tier cloud application. At the same time, system counters were monitored to verify the system load. Finally, we present a model that allows estimating the response time based on historical information about the load on the operating system, and some evaluations of this model.

This paper is organized as follows: Section II introduces a background on response time and performance counters of operating systems; Section III presents related work; Section IV describes preliminary experiments used to fit a new model; Section V presents evaluations; finally, Section VI concludes the paper and addresses future work.

II. BACKGROUND

QoE [8] refers to the user's perception of the quality of services transactions and may be represented by human feelings. On the other hand, QoS refers to a systematic method to evaluate the service, usually via metrics that can be measured and verified and that directly affect the perception of the end user. A QoS metric that directly impacts the QoE is the response time of applications. This section presents a conceptualization of response time and shows some involved algorithms for its calculation.

A. Response Time

Response time can be understood and evaluated in several ways depending on the context. In this section, we will conceptualize and explain the response time in the context of a cloud computing infrastructure. Figure 1 presents a three-tier architecture, in which a customer performs requests to a service in a cloud, and waits for a response.

The time required to complete the entire process between the start of the request from a customer up to all of its

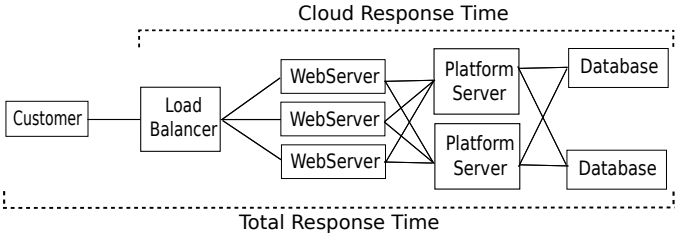


Figure 1. Response Time in Cloud Architectures

response consisting of the total response time as shown in Figure 1. Thus, the total response time includes the time that packets must travel between the customer and the cloud. This means that the cloud manager that is providing the service has no way to measure and ensure the quality of service of this external link, unless there is monitoring from the client. However, a monitoring client-side influences other aspects such as security, privacy, and the actual cost of monitoring on the link.

Therefore, when dealing with response time in cloud infrastructure, we are assuming the Cloud response time. The time spent in establishing the connection operations can be seen in (1), where one-way trip (OTT) time is the difference between the last (ω) synchronization packet (SYN) with the first (α) acknowledgment packet (ACK), divided by the number of participants in the connection.

$$OTT = \frac{\alpha ACK - \omega SYN}{2} \quad (1)$$

Therefore, the Total Real Time (RT) (Figure 2) of each request is the sum of the one-way trip times, and the time that a reply takes to answer a client's request.

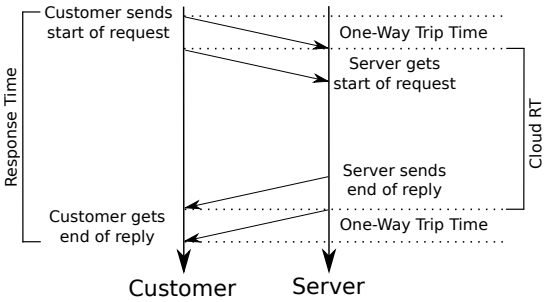


Figure 2. Total Response Time and Cloud Response Time

Figure 2 shows the one-way trip in the beginning and end of the connection establishment, and therefore, the transmission between the client and the server.

In summary, the cloud response time consists of the time that the customer request takes to be processed by the cloud service provider. This is counted from the moment it is received by the cloud application layer, through the business layer, searching and querying data in the database layer, and

 ISBN: 978-1-61208-398-8

returning necessary information for the application layer to be sent to the customer.

B. Operating System Counters

The response time is an ideal metric to verify the quality of a service offered via the network, but its monitoring may cause overhead. Estimate the response time of an application without interference in the channel between the customer and the service can substantially reduce this impact. To this end, we hypothesized that operating system counters can be monitored in order to verify the response time within the cloud data center, and therefore, allow estimating the response time between the customer and the data center is acceptable or not.

In the context of this work, we target real-time Linux counters as the focus of monitoring due to the fact that Linux is widely used in IaaS environments, such as Openstack [9]. Linux stores the counters in a virtual file system referenced in `/proc`. This directory contains, rather than files, a runtime information system in which one can monitor the states and loads of the processor, memory, and devices, in real-time. In particular, we rely on information available at the hypervisor level (rather than virtual machine-level) when the hypervisor is supported by a privileged operating system, which is the case of Xen [10] (on its Dom0).

Because of this, most system management commands search for system information into this directory. Since this information is accessible at the user level, system management tools can display mashups that will support decisions about the use of resources to the system administrator, e.g., system and services [11].

To monitor the load average of I/O in our tests, we used the information provided by `/proc/loadavg`. This file is populated with values collected from the run queue of the operating system. It stores a series of values of load in three intervals representing the average load of the system in the last 1, 5, and 15 minutes. These values are updated by the system every one minute. Since the values are shown in the time period, it provides a good indication on if the workload is increasing or decreasing the use of resources. Moreover, it allows to estimate when the system is overloaded and impacting on QoS metrics.

III. Related Work

Aceto et al. [12] discusses the difficulty of monitoring cloud environments with respect to the mapping of high-level metrics to metrics at the infrastructure level, due to the fact that high-level metrics may include external parameters to the infrastructure, which are not controlled by cloud environment.

Emeakaroha et al. [6] proposed a framework for managing the mapping of low-level resource metrics to high-level SLAs. The scalability of the model was validated using queuing network models, and it was able to detect SLA violations and notify the manager module of the cloud environment.

Dobson and Sanchez-Macian [13] presented a work-in-progress paper proposing a QoS ontology that can be applied on several scenarios of cloud/grid. This model is divided into two parts: QoS monitoring and QoS adaptation.

Rosenberg et al. [14] discussed QoS attributes for web services, identifying the most important attributes and its composition from resource metrics. In addition, the study presented some mapping techniques to compose QoS attributes of resources to generate metrics of SLA parameters for a specific domain.

D'Ambrogio and Bocciarelli [15] present a model-driven approach with the intention to incorporate application performance prediction into a service composition process. The paper shows the composition of SLA parameters, although it does not consider monitoring the SLA.

Comuzzi et al. [16] propose an architecture for monitoring SLAs considering two criteria: the availability of historical data to evaluate the SLA offers and the evaluation of the ability to monitor an offer of SLA.

What differentiates our work from the other ones discussed in this section is the fact that we present a methodology to estimate the response time without impacting on user communication channel, either in terms of performance or in terms of security and privacy. To the best of our knowledge, no other work has mapped information from operating system counters, obtained from the hypervisor, in order to infer the performance of an application running in a virtual machine that directly influences end users' QoE.

IV. Estimating Application Performance

For web applications, particularly in the case of e-commerce, performance tests are essential. A service provider can not define the needed amount of resources required by the application to serve a specific workload based only on average traffic. To ensure that a web application meets certain criteria such as performance, data throughput or response time, test in an environment similar to the production environment is required.

The focus of this paper is on three-tier cloud applications, type of most used application on cloud environments. Therefore, we are targetting a virtualized environment that supports a web interface that accesses, through a business logic layer, a database allocated to another computer system. This architecture is shown in Figure 3.

The first step toward the goal of performance estimation was conducting experiments to enable the modeling of the relationship between response time and the load as indicated by the loadavg counter, which we detail next.

 ISBN: 978-1-61208-398-8

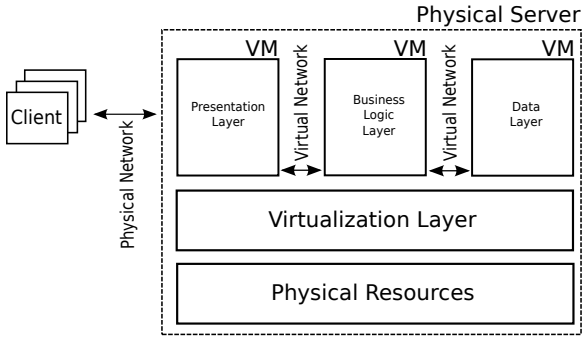


Figure 3. Three-Tier Cloud Architecture

To this end, we deployed a testbed consisting of two physical servers connected by a Gigabit Ethernet network. Each server has two Intel Xeon 2.4 GHz processors (12 cores) and 16 GB RAM. On each server, we deployed three virtual machines, each one supporting the web server (Apache), application server (Tomcat) and database server (MySQL). The CPUs are dedicated for each VM and disks are shared between them. Load average is taken from /proc/loadavg file at the hypervisor level (Dom0).

To represent a multi-layer architecture, the Apache Benchmark was used. This benchmark is widely used to test performance of multi-tier applications. It mimics the users' access to web servers serving as front end for multi-tier applications. Aiming to emulate an elastic cloud environment, we use the HAProxy load balancer that splits the workload between new nodes, when the response time set out in a SLA is exceeded. In our tests, 100 clients should insert 1000 records into a table, query them, and delete them at a maximum of 300 milliseconds using 50% of the network throughput. To control the flow of the network, we use the Linux Traffic Control (TC). This allows adjustment of the network flow and hence impact the response time of applications.

The evaluations were conducted in a total of 45 minutes (2700 seconds). Each test was performed with the size of 15 minutes because this is the maximum time that the loadavg uses to update its values. For the tested architecture, the loadavg values can range from 0 (underused) to 22 (overutilized). We set 300 milliseconds as the target response time of the test application. The values of loadavg and their corresponding timestamps were recorded in intervals of 1 second. The application response time was monitored by a feature of the Apache server, which shows when a request is received from a customer, and when it was responded to the customer, which allows the verification of the infrastructure response time. As Apache requests are also based on timestamps, it was possible to relate the load as measured by loadavg with the request response time as measured by Apache. The experiment was repeated 35 times and the average values obtained on each experiment time are reported. As the results showed little variance, 35 experiment repetitions allowed statistically significant analysis of results. In each repetition, the stochastic process was run with a different seed value. The hypothesis of correlation between the two measurements was tested using the Pearson correlation coefficient

A. Evaluation and Discussion

Figure 4 presents the tests results as the average values collected by loadavg and Apache. The values presented a small standard deviation, and thus the deviation is not shown. During the first 15 minutes (900 seconds), the SLA response time of 300 milliseconds was met in its entirety. In the next 15 minutes, the network latency has been increased to enable us to evaluate the effect of resource underutilization by forcing requests to arrive at a lower rate. Similarly, in the last 15 minutes the network latency is reduced to induce a higher request income on the system, causing a longer response time.

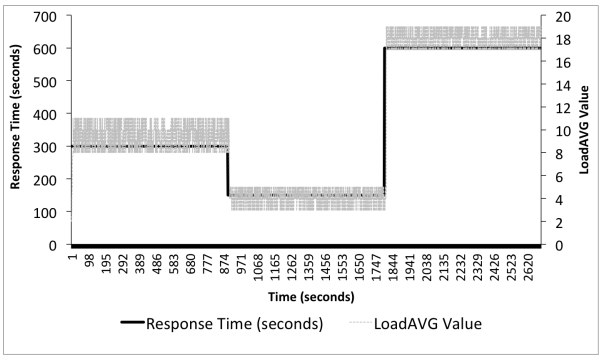


Figure 4. Response Time vs. LoadAVG Evaluation

Figure 5 shows that the load values given by loadavg accompany the application response time behavior. The Pearson correlation coefficient between the behavior of the application response time and load values of the operating system displayed by loadavg was 0.9965. This shows that there is a strong positive correlation between the two behaviors.

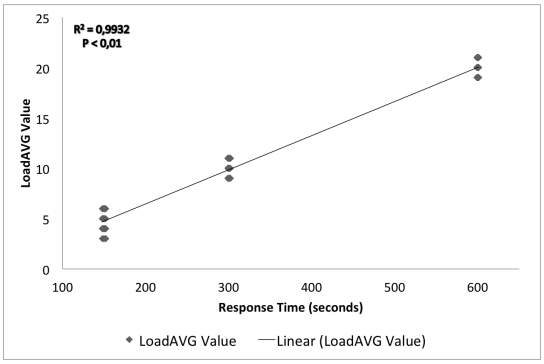


Figure 5. Correlation Plot between Response Time and LoadAVG Values

Based on these results, we can estimate, from the infrastructure, how much the response time is being influenced, and if there is the need for more or less resources so that an SLA is met. The next step towards enabling the estimation of the application response time using loadavg was building a model that captures the relation between these two variables. The modeling process is discussed next.

B. Modeling

Because loadavg values are stored for logging purposes over time, these values can be analyzed to verify that the environment is going towards overutilization or underutilization. This data can be used to automatically scale the application resources to an amount that meets application needs.

The correlation plot from Figure 5 provided some evidence that a linear regression model would be a good fit for our model to estimate the response time. The purpose of multiple regression is to predict or estimate a dependent variable y, in response to the values taken by a set of independent variables X. In this case, we assume our dependent variable Y to be the response time, and we estimate it using the value of loadavg from the last 1 minute, 5 minutes, and 15 minutes (our independent variables).

$$Rt = c_0 + c_1 \cdot avg1 + c_2 \cdot avg5 + c_3 \cdot avg15, \qquad (2)$$

Where Rt represents the estimated cloud response time of applications in the entire nodes. The coefficients c_0, c_1 and c_2 are the weights assigned to each variable, in each loadavg times. $avg1$, $avg5$, and $avg15$ are monitored available values in `/proc/loadavg`, for 1, 5, and 15 minutes, respectively. The c_0, c_1, c_2, and c_3 (4.133, 1.005, 1.478, and 1.597, respectively) values serve to the best fit of the model, and show the curve in the correlation plot.

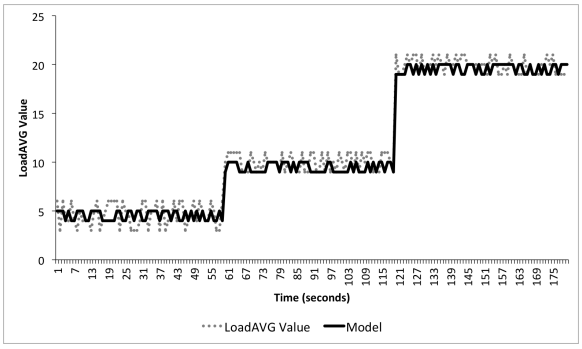


Figure 6. Model Accuracy Test

An important validation aspect of the model is the residue analysis, which shows the model significance and evaluates the contributions of regression variables. In the proposed model, it is possible to assert that all the points follow the behavior of the line, indicating that the errors are normally distributed.

V. Model Evaluation

This section presents an evaluation of our model applied in a real environment, aiming to validate the proposed model.

A. Testbed

For conducting the evaluation of our model, we used HammerDB, because it is a tool that reproduces the behavior of most applications offered by cloud environments as a service. It supports multiple clients accessing a three-tier service. Another advantage of HammerDB is its flexibility for tests configuration. Both the number of users and the number of simultaneous connections can be set and changed on-the-fly during the test. This allows a script to be created to describe the desired behavior. The client-server environment was deployed on the same two physical servers used in preliminary experiments, and using the same network between them.

As the HammerDB allows controlling the client application behavior (increasing or decreasing the number of users and connections), we choose three behaviors that represent routine situations in environments that support cloud services: The first scenario represents an environment in which the arrival rate of requests is low, and gradually increases over the time; The second scenario represents an environment in which initially the arrival rate of requests is low, and gradually increases over the time, and then returns to a low rate; The third test represents an arbitrary behavior fluctuating between moments of slow and high rates of requests.

Each test was performed for 30 minutes, and the results of the real environment tests and the results generated by the model based on loadavg data were compared. All tests were performed 35 times, and the repetitions did not show a significant variation in the results less than 0.03%.

B. Results

Figure 7 depicts the results for the first scenario, where the increase in application demand causes the response time to worsen until the end of the test. We can see that the behavior of the proposed model follows the real trace behavior. Importantly, we can notice that the estimating response time presents variations between real and model results. This difference is due to loadavg update times, which features a refresh rate of 1 minute. Thus, the model responds at least at 1 minute later on the real response time. This update time is inherent to the Linux kernel, and perhaps in future versions, if that time is reduced, the model can track more quickly the application response time.

 ISBN: 978-1-61208-398-8

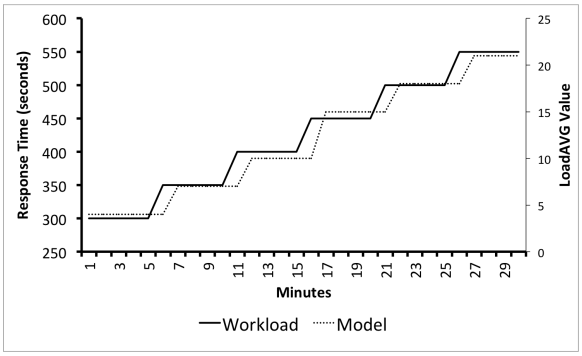


Figure 7. Scenario starting from a situation of an acceptable rate to a gradual increase in the number of requests

Although there is a delay in the model in relation to the real response time, it remains faithful to the real trace, as we can see in Figure 8. Even when the response time worsens and returns to an acceptable level, the model can track all changes and represent, with a very little difference, the real behavior of the response time.

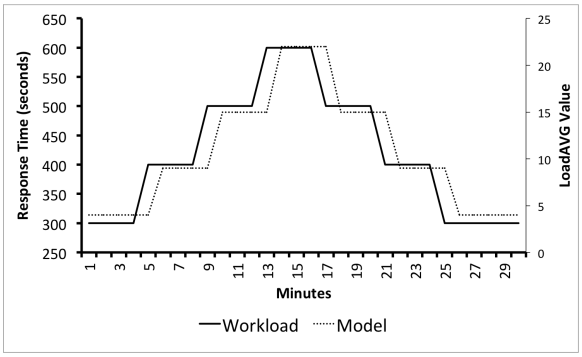


Figure 8. Scenario with fluctuation in the number of requests following a normal distribution

The last test is shown in the Figure 9, where the response time behavior is very diverse, often fluctuating and forcing the model to fit faster way to these changes. Still, the model could represent the fluctuations of the response time fairly well.

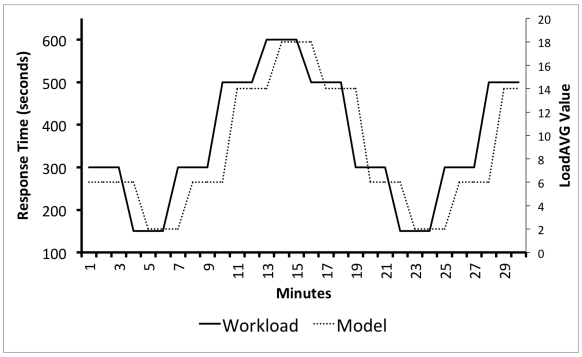


Figure 9. Scenario with arbitrary fluctuation in the number of requests

The model proved to be adjusted during the execution of the three tests against real environments (Figures 7, 8, and 9). To confirm the trend of the model results against the real tests, again we used the Pearson's correlation coefficient. Thus, we can see in Table I, the results for each set of tests. The results showed that on average, the model is accurate to represent the cloud response time up to 94%. The three tests showed P-value < 0.01, what enables us to reject the null hypothesis that both variables are independents.

TABLE I. PEARSON CORRELATIONOF THE TESTS

Test Scenario	*Correlation Coefficient*	*P-Value*
1)	0.944	$p < 0.01$
2)	0.968	$p < 0.01$
3)	0.932	$p < 0.01$

C. Limitations

Models, in most cases, present some limitations to represent real production environments. Our model presents limitations about the accuracy during runtime, because the load values show a delay when compared to the real executions. This is not exactly a limitation of the model, but an operating system counter feature that displays a delay of 1 minute between each update. Unlike High Performance Computing (HPC) jobs that must have a finite time, cloud services mostly keep running throughout the service run time. Thus, cloud services do not have a time of total execution time. So the operating system counter update time (and model) can be neglected, because one minute will not impact strongly on the final result. In addition, the update time is a choice by kernel architects, and this time can be reduced in the future, either on a future release of the Linux kernel. Furthermore, as Linux is an Open Source software, its source code can be easily modified and a new kernel generated that provides more frequent updates in the loadavg.

The second important point to note consists of the architecture used to feed and create the model. The preliminary and final tests were based on one hardware/software architecture. However, the model allows weights adjustment in their c_1, c_2 and c_3 coefficients, which will allow the fitting for each new environment.

The most important feature of the model is to estimate, at one point, what is the application response time. With this information at hand, the infrastructure manager can make decisions in order to maintain an acceptable response time in case of overuse, through the new entities deploy in the resource pool, or in the case of underutilisation, save costs with removal of these. However, although there are limitations set forth above, the main advantage of this model consists of not requiring access to the virtual machine nor interfering directly in the channel between the client and the server, which could cause overhead during a measurement process, slowing the channel.

 ISBN: 978-1-61208-398-8

We believe that our methodology applies, because in cases of flash crowd, the excess requests do not cause excessive utilization of resources in the same proportion: the metrics analyzed in our work displaying the machine's point of view, then the utilization of resources can not go beyond 100%. In this case the excess requests are rejected.

VI. Conclusion and Future Work

Cloud computing environments are rapidly becoming a standard platform to support computational services. As the access to these services is performed via the network, this becomes a decisive factor for the quality of services offered. Among the performance metrics and quality involving quality of users' experience of services, one of the most important is the response time. In addition to impact on the user experience, the response time could indicate that available resources may be insufficient to ensure the proper functioning of the service to users, impacting both users' cost and the total cost of ownership.

High-level metrics, such as response time is often measured between the customer and the cloud provider. However, the infrastructure that supports cloud environments (IaaS) typically monitors low-level metrics, such as CPU, memory, and network usage. Therefore, the translation between high-level metrics to low-level metrics is a very complex challenge in today's cloud environments. Moreover, monitoring high-level metrics are quite difficult, and impact performance and privacy issues of the user. Furthermore, the communication channel between the customer and the cloud listener cannot be monitored by the cloud infrastructure without overhead on the communication channel.

In this paper, we proposed the use of loadavg, a counter of the operating system that stores information on-the-fly on the load of the node, as a parameter to estimate the behavior of the response time. Based on this, we developed a model that analyzes the loadavg values, and estimates what the response time of applications with 94% of accuracy. Such model allows to estimate the time to process the cloud application request within the infrastructure and without any impact on the users or their communication channel. As a future work, we intend to explore other operating system counters such as iostat and netstat, to develop more complete models.

References

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computing Systems, vol. 25, no. 6, Jun. 2009, pp. 599–616.

[2] K. Kritikos, B. Pernici, P. Plebani, C. Cappiello, M. Comuzzi, S. Benrernou, I. Brandic, A. Kertész, M. Parkin, and M. Carro, "A survey on service quality description," ACM Computing Surveys, vol. 46, no. 1, Jul. 2013, pp. 1:1–1:58.

[3] W. Tang, M. S. Kim, and E.-N. Huh, "Analysis of qoe guarantee on hybrid remote display protocol for mobile thin client computing," in Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication, ser. ICUIMC '13. New York, NY, USA: ACM, 2013, pp. 118:1–118:8.

[4] V. Gabale, P. Dutta, R. Kokku, and S. Kalyanaraman, "Insite: Qoe-aware video delivery from cloud data centers," in Proceedings of the 2012 IEEE 20th International Workshop on Quality of Service, ser. IWQoS '12. Piscataway, NJ, USA: IEEE Press, 2012, pp. 8:1–8:9.

[5] P. C. Brebner, "Is your cloud elastic enough?: Performance modelling the elasticity of infrastructure as a service (iaas) cloud applications," in Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering, ser. ICPE '12. New York, NY, USA: ACM, 2012, pp. 263–266.

[6] V. Emeakaroha, I. Brandic, M. Maurer, and S. Dustdar, "Low level metrics to high level SLAs - LoM2HiS framework: Bridging the gap between monitored metrics and SLA parameters in cloud environments," in International Conference on High Performance Computing and Simulation (HPCS), 2010, June 2010, pp. 48–54.

[7] V. C. Emeakaroha, T. C. Ferreto, M. A. S. Netto, I. Brandic, and C. A. F. De Rose, "CASViD: Application level monitoring for SLA violation detection in clouds," in Proceedings of the 2012 IEEE 36th Annual Computer Software and Applications Conference, ser. COMPSAC '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 499–508.

[8] R. Schatz, T. Ho, L. Janowski, and S. Egger, "Datatraffic monitoring and analysis," E. Biersack, C. Callegari, and M. Matijasevic, Eds. Berlin, Heidelberg: Springer-Verlag, 2013, ch. From Packets to People: Quality of Experience As a New Measurement Challenge, pp. 219–263.

[9] T. Rosado and J. Bernardino, "An overview of openstack architecture," in Proceedings of the 18th International Database Engineering & Applications Symposium, ser. IDEAS '14. New York, NY, USA: ACM, 2014, pp. 366–367.

[10] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," SIGOPS Operating Systems Review, vol. 37, no. 5, Oct. 2003, pp. 164–177.

[11] D. Josephsen, Building a Monitoring Infrastructure with Nagios. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2007.

[12] G. Aceto, A. Botta, W. De Donato, and A. Pescapè, "Cloud monitoring: A survey," Computer Networks, vol. 57, no. 9, Jun. 2013, pp. 2093–2115.

[13] G. Dobson and A. Sanchez-Macian, "Towards unified qos/sla ontologies," in Proceedings of the IEEE Services Computing Workshops, ser. SCW '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 169–174.

[14] F. Rosenberg, C. Platzer, and S. Dustdar, "Bootstrapping performance and dependability attributes ofweb services," in Proceedings of the IEEE International Conference on Web Services, ser. ICWS '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 205–212.

 ISBN: 978-1-61208-398-8

[15] A. D'Ambrogio and P. Bocciarelli, "A model-driven approach to describe and predict the performance of composite services," in Proceedings of the 6th International Workshop on Software and Performance, ser. WOSP '07. New York, NY, USA: ACM, 2007, pp. 78–89.

[16] M. Comuzzi, C. Kotsokalis, G. Spanoudakis, and R. Yahyapour, "Establishing and monitoring slas in complex service based systems," in Proceedings of the 2009 IEEE International Conference on Web Services, ser. ICWS '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 783–790.

 ISBN: 978-1-61208-398-8

Control Plane Routing Protocol for the Entity Title Architecture: Design and Specification

Natal Vieira de Souza Neto,
Flávio de Oliveira Silva
and Pedro Frosi Rosa

Faculty of Computing
Federal University of Uberlândia
Uberlândia, MG, Brazil
Email: natal@mestrado.ufu.br, flavio@ufu.br,
pfrosi@ufu.br

João Henrique de Souza Pereira

Innovation, Research and Development
Algar Telecom
Uberlândia, MG, Brazil
Email: joaohs@algartelecom.com.br

***Abstract*—Current and future applications pose new requirements that Internet architecture is not able to satisfy. In this context, new network architectures, focusing on different aspects, are being designed and deployed. The Entity Title Architecture (ETArch) is a clean-slate Software Defined Networking based approach which aims to satisfy different applications requirements such as multicast traffic, mobility and Quality of Service. This work presents the design and specification of the routing protocol used by ETArch, by describing the services, primitives and associated rules. This work contributes with ETArch in a central point of the architecture, the inter-networking. The multi-objective routing mechanism described in this work takes into account, applications requirements such as mobility and security. Moreover, the protocol presented works at the control plane and uses ETArch workspace concept, representing a new class of routing mechanism that differs from classical approaches such as Distance Vector or Link-State.**

Keywords–Software Defined Networking; Routing; Protocol Specification.

I. INTRODUCTION

The advances in software, hardware and communications brought a world where mobile devices with high resolution cameras, different sensors, equipped with multiple wireless interfaces are connected to clouds of servers using broadband access networks. New services and applications emerged and the Internet architecture [1], proposed in the sixties, that also collaborated to this scenario, is not able to satisfy the new applications requirements such as mobility, Quality of Service (QoS) and security [2].

To face these challenges research groups around the world [3][4] are designing and deploying new network architectures, some of them, based on a clean-slate approach in order to bring life the Future Internet.

One of these network architectures, which uses a clean-slate approach, is the Entity Title Architecture (ETArch), based on the Entity Title Model [5]. ETArch uses a naming and addressing scheme based on a topology-independent designation that uniquely identifies an entity, named Title, and on the definition of a logical bus which gathers multiple entities, willing to communicate driven by specific purpose, named Workspace. Workspaces are, dynamically, created/removed according to user's specific needs. Users could be linked (attached) to the workspaces during its life cycle.

The workspace is capable to handling the requirements of users and applications over time. Our research regarding ETArch demonstrated some of these requirements such as naming and addressing [5], QoS [6], multicast [7] and mobility in [8].

A central point in an architecture are the inter-networking mechanisms, which allows the network to scale its coverage in a geographical perspective. This work contributes with ETArch, by presenting the design and specification of the routing protocol. To do so, it describes the services, primitives and rules associated to this protocol. This paper is a conceptual work to define the routing mechanism to ETArch, that runs in the control plane by using a Software Defined Networking (SDN) as its underlying interconnection strategy.

The remaining of the paper is structured as follows: Section II presents an overview of related work about routing on the Internet and new network architectures. Section III introduces ETArch concepts. Section IV describes the routing approach and mechanisms defined to ETArch, and finally, Section V presents some concluding remarks and future work.

II. RELATED WORK

Several algorithms like Chandy-Misra, Merlin-Segall, Toueg or Frederickson were proposed to handle routing [9]. Also, according to [9], none of these approaches is good to be used in large scale networks, such as the Internet. In the Internet architecture, two classic approaches were more used: the link state routing and the distance vector routing. Several other algorithms use both approaches or one inspired approach [10].

The approaches adopted by the Internet architecture are based on a weight on the graph edges. The weight is calculated using the distance through number of hops or queuing time [10]. The main difference between both is based on the knowledge of complete network topology or only the directly connected neighbors, as used by the distance vector.

As long as the Internet scaled, different researchers focused on the routing problem by considering the new scales and requirements. One of the proposed approaches is related with

 ISBN: 978-1-61208-398-8

the separation of the routing into two different layers using IP addresses [11] or even by rethinking the network layer routing and forwarding [12].

Another approach presented by the research community is the aggregation, where a router aggregates others, and a hierarchical vision is possible with a router being the parent of others. For the use in scalable problems, the aggregation is a interesting approach. The compact routing was also proposed to resolve problems related to big number of data in the routers, where the lines in the routing table grow fast [13][14].

Yang proposed the New Internet Routing Architecture (NIRA) [15] which gives users the ability to choose different sequence of providers in order to routing the primitives. So the users can make this choice taking into account the type of applications. The work proposed here also has this approach in common; however, the routing is affected by the applications requirements considering QoS and also Quality of Experience (QoE) parameters. Besides that, the primitives could be routing using different paths simultaneously or by having an alternative path on its header, that could be used in case of a failure of the initial one [16][17].

On the sensor networks, some protocols have been proposed for the secure routing, by forwarding every network node and creating routing tables by using grouping algorithms [18]. Also, the scaling problems of the Internet were treated in some works, and the routing is presented, and the concern relative to the routing in Future Internet [19].

The works cited present different forms to treat the routing problems at the TCP/IP architecture and other architectures proposed in some works. Also, some architectures focused on Future Internet was proposed like Recursive InterNetworking Architecture (RINA) [20] and ETArch [5]. The latter uses a horizontal address and does not have some TCP/IP characteristics, by separating identifier and localizer. With ETArch, it is possible to apply SDN concepts to use other communications approach prepared to new challenges on the future [21].

The workspace in ETArch is described in the next section. The concept of workspace is similar to other concepts introduced by the community. The Open Network Operating System (ONOS) [22][23] project is using the Intent concept. A similarity of workspace and Intent functions can be found.

ETArch is similar to other projects that treat clean-slate Routing, but the main difference of this work is to propose a routing protocol processing only in the control plane. An example of other project is Mobility First, also a clean-slate approach, but it treats routing using the traditional way, with processing on control and data planes [24][25]. So, the ETArch novelty is the clean-slate architecture able to meet the application requirements.

III. ENTITY TITLE ARCHITECTURE

The creation of the Entity Title Architecture was motivated by some requirements found in the current Internet: Energy Efficiency, Mobility, Multicast, Quality of Service (QoS), Scalable, Security, and others. The problem of routing is associated with several requirements. The architecture concepts were presented on [7][8][21][26].

The routing made in the current Internet architecture considers basically the distance between two hosts, and the packets being transferred knows the final destination (identified by the host). This is problematic because the hosts, which is involved in the communication, should be identified using a hierarchical structure. Currently, this identification is made setting an Internet Protocol (IP) address to each host. When a host changes its location, the communication being done is lost, because the identification changed. The idea of set IP therefore is questionable, and no set IP is a problem considering that the current algorithms use IP to transmit the packets from source to destination.

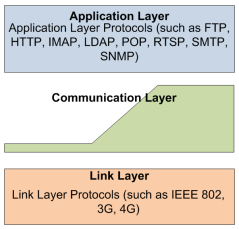


Figure 1. ETArch Layer Architectural Pattern

Another remaining problem of the current Internet is the multicast traffic. Today, a communication is made between two hosts, then all packets contain the source and destination IP addresses. A packet sent from the source host does not have a list of destination hosts, but only one IP address relative to one host. It is clear that the routing in new architectures can not to consider the destination host, but what is being located. The separation between identifier and location of a host is present in several Internet protocols by regarding for the Future Internet. A similar concept are the Content Distributed Networks (CDN), where the initial look for a destination is not made looking for a host, but a content, independently of its location.

Energy efficiency and QoS are two requirements to future routing. The new forms of routing to Internet should be able to consider these requirements, and not only best way based on distance between two hosts. It is necessary because one application can order routes that consume less energy or that have different level of QoS.

As mentioned at the Introduction, ETArch does not have a fixed layer like TCP/IP structure. A communication layer was introduced where TCP/IP would be. This logical layer is different of the TCP/IP layer, because as shown in Figure 1, it is flexible depending of the requirements in specific communications. An example is an application that ordering only one requirement (like determined bandwidth, for instance), then the communication layer need to prepare the scenario considering this bandwidth. In another example, an application may request several requirements (like bandwidth, low consumption of energy and QoS, for instance), then the communication layer need to prepare the scenario considering all these requirements.

To create this flexible layer, able to provide communication with different requirements, the ETArch was designed using some concepts, summarized below:

- Entity: Anything that can communicate. So, an Entity can be a host, a user, a process, a link, a mobile phone, and others. Independently if it is responsible to run applications or to control the network.
- Title: it is an unambiguous name to uniquely identify an entity. For example, a workspace has a Title.
- DTS: The Domain Title Service is responsible to control the network. It works like a controller present in SDN, handling the communication and the requirements over time. It is composed by agents named DTS Agent (DTSA) which act as controller.

 ISBN: 978-1-61208-398-8

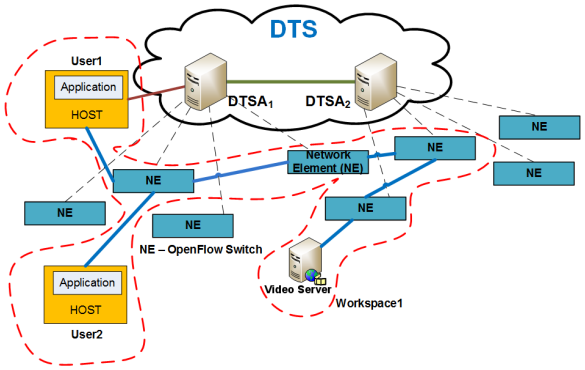


Figure 2. ETArch Workspace Vision

- Workspace: it is a logical bus shared by different entities, where an entity sends and receives information to/from other entities.

The communication in ETArch is made by using the workspace. Workspaces do not exist in initial network setup. If an entity wishes to transmit/provide something, it must create a workspace, which is registered by the DTS through their agents (DTSA). Whether another entity wishes to share something provided by one previously created workspace, it must be linked (attached) with this workspace. This way, all entities which wish to participate to the transmission must be attached on the workspace. The DTS is responsible to control the various workspaces that entities can create and attach. Capabilities are assigned to the workspace such as bandwidth, and so on.

The DTSA has information about the topology of the Network Elements (NE) in its domain. Current works over ETArch use OpenFlow switches as NE, and each NE is controlled by a DTSA. All workspaces created by the entities are stored by the DTS workspace database, then, when an Entity wants to attach in a workspace, the DTSA has enough information about the specified workspace.

The first problem to be treated in ETArch is when a given DTSA does not have information about a workspace, which has been registered in another DTSA. Thus, the routing protocol proposed in this paper starts by the problem of finding a workspace which is being seeked by a DTSA.

Figure 2 shows an example of an existing workspace. In the figure, two DTS Agents ($DTSA_1$ and $DTSA_2$) are controlling some NE. Note that each DTSA has topology information about all NE in its domain. A Video Server is connected to a NE controlled by $DSTA_2$. The Video Server is an entity and it is streaming a video using $Workspace_1$. When the entity named $User_1$ decides to join (attach) in $Workspace_1$, and when it occurs all information sent by the Video Server will be forwarded to $User_1$. In implementations on the ETArch, when the information about the attaching of $User_1$ with $Workspace_1$ arrives on the DTS (control plane), the two DTSAs configure rules on NE to forwarding all data of $Worskapce_1$ to specific ports, so that data goes to the $User_1$. At this moment, the data plane has only one consumer entity and it seems an unicast connection. If the entity $User_2$ decides to join (attach) to $Workspace_1$ the same behavior is made, and the DTSAs configure rules to forwarding the data to User1 and User2. Look that the traffic is split in the NE around the consumer entities.

Thus, at ETArch, the problem on routing has an additional problem and it is divided in two parts: i) how to find a workspace, as it is not fixed as hosts in the Internet; and, ii) how to include the Entity requirements in the best route choice.

Current works on ETArch and their implementations use the distance to define de best route. It works, however, to continue evolving the architecture is necessary a routing protocol prepared to use the requirements required by the applications. An Entity can order low energy consumption, a minimal bandwidth or specific QoS rule. So, the routing protocol proposed on the next section is prepared to work with requirements in different levels. A level can be low, medium or high, ie, the Entity can order a requirement like low distance and high energy consumption: it means that the best route to this Entity should be a low energy consumption (because is a high level for the Entity), and the distance need not be the lowest considering the topology. Between low, medium and high levels, the DTSA decides the best path to the workspace be expanded to the Entity.

A workspace allows the DTS knows the requirements of the applications and adjust network to satisfy these requirements. In current Internet, the application layer is responsible to treat requirements, and take it into the network is the main contribution when using workspace concept.

ETArch currently does not provide communication between different DTSAs, and the routing protocol proposed on the next section is prepared to provide an Entity joins in a workspace in other DTSA. The problem with two DSTA is how the DTSA searchs a workspace that is not in its local database.

IV. Routing Algorithm Specification

As can be seen in Equation 1, the processing time (Tp) of a router can be considered the sum of routing time (Tr) and the sum of switching time (Ts), to all n packets. It means that each data in a router has the processing to determine the route and the switch times of the packet.

$$Tp = \sum_{i=1}^{n} Tr_i + \sum_{i=1}^{n} Ts_i \quad (1)$$

The idea of the new approach proposed here is to the Tp in the data plane is only the sum of Ts. In this way, the sum of Tr must be zero. Since the entire route can be defined on control plane, before the communication starts, ETArch permits to use the SDN concepts to the data plane be only responsible for switching.

In ETArch, the workspace concept permits routing considering several requirements, like number of hops, bandwidth, energy consumption and others. In the architecture, the Entity calls the DTSA asking by a given workspace, and the DTSA looks for the workspace and extends it, through several NE to reach the Entity.

The idea proposed in this paper is to use the workspace concept to reach the destination. When an entity wants receive some data it must look for a workspace. Then, it sends to DTS (the controller) the request for the workspace attach. The DTS attaches the entity on it, and the transmissions over the workspace are sent to the entity.

Previous works [5][26] explain how these rules are configured in the NE to extend a workspace through the network.

 ISBN: 978-1-61208-398-8

Here is proposed a protocol design to find a workspace by considering DTSA topology and the link of several DTSA.

The routing protocol runs over control plane. It was built considering two situations: when a workspace is in the same DTSA as the Entity that wishes attach; and when a workspace is not in the same DTSA. The first situation is named intra-DTSA routing and the second is named inter-DTSA, similar to inter-networking or inter-domain used in current networks and topologies.

A. Intra-DTSA Routing

The intra-domain routing happens when an Entity requests an attach with a workspace and the DTSA which controls the Entity has enough information about the specified workspace. This information can be: the workspace was created by an Entity plugged in that DTSA; or some Entity in that DTSA is already attached with the workspace. Then the intra-DTSA needs only to extend the workspace, i.e., to inform all NE on the way the new rules to forwarding data to the Entity.

The DTSA must be prepared to recognize Entity requirements and to decide the best path. It can be stated that the intra-DTSA routing is similar to the link state mechanism.

B. Inter-DTSA Routing

The inter-DTSA routing happens when an Entity requests an attach with a given workspace and the DTSA, which controls the Entity, has no information about the specified workspace. In this case, the first step is to look for the workspace.

For this propose, the first service to be specified is WORKSPACE_LOOKUP. This message is sent from a DTSA to its Master. A Master DTSA is the resolver of several DTSAs and has information about all workspaces created or extended by any Entity.

The Figure 3 shows some DTSAs linked with a Master-DTSA (D_1). The control informations sent by the DTSA, as WORKSPACE_LOOKUP, for instance, are sent through a control workspace. This is a private control workspace that creates a bus between some DTSAs and their Master. When a DTSA receives the attach information from an Entity and it does not have information about the workspace, it sends a WORKSAPACE_LOOKUP to the Master-DTSA.

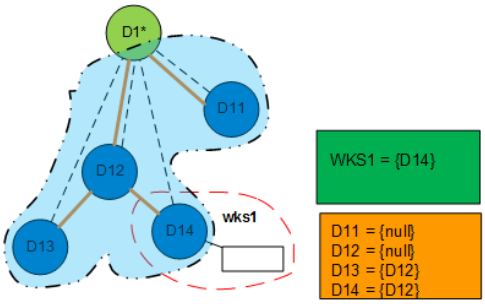


Figure 3. Private Control Workspace

The protocol has an algorithm to decide the best route. It starts when a WORKSPACE_ATTACH message arrives in the DTSA. The Master-DTSA has information about all workspaces extended or created by any NE in its domain. When a workspace is requested and the Master does not have knowledge about the workspace, it should forward the message to other Master DTSAs. A Master DTSA has a control workspace with other Master DTSAs as shown in the Figure 4, where the Masters D_1 and C_1 are connected by a control workspace. This control workspace is named Public and two or more Master-DTSAs are linked by this workspace, and the level of this linking can be worldwide.

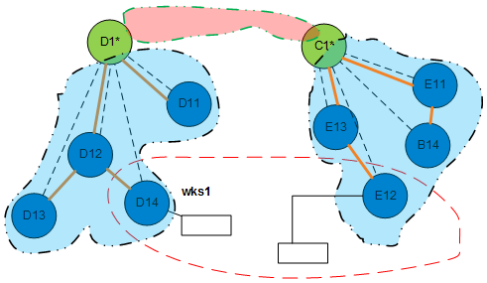


Figure 4. Public Control Workspace

It is important to note that there are two timeouts when WORKSPACE_LOOKUP is sent. The first is the max time to the DTSA to receive the first path. The second is the max time that the DTSA waits more paths when it already received the first path. It is because a request can result zero, one or more responses.

When the DTSA receives responses, containing a list of paths to extend the workspace, it must decide the best path considering the Entity requirements. Thus, a list of paths works as a routing arguments, being that each member of the list contains paths and information about capabilities. Upon deciding the best route, the DTSA must send a WORKSPACE_CONFIGURATION to each DTSA in the path. This service contains information to each DTSA in the path to extend the workspace until the Entity. When a DTSA receives a WORKSPACE_CONFIGURATION it must: analyzes its topology and decides a route in its NE to extend the workspace. After this, it must send to each NE on the path, rules containing workspace name and new port of out. After this the workspace is extended by that DTSA and the DTSA inserts this information in its local database for next requests does not need new WORKSPACE_LOOKUP.

The execution order of the protocol algorithm in the DTSA is presented in Figure 5. The functions invoked in pseudo code are executed by the DTSA itself, and they can modify the object wksInfo, which maintain informations about the workspace and routes. Also, in the procedure onReceive-WorkspaceAttach, some functions need of specific service messages on the routing protocol, like notifyMaster and requestConfiguration.

The notifyMaster function exists because all workspace extended in some DTSA must be notified to the Master. Thereby, the Master-DTSA can insert in its local database and, when a lookup arrives in Master, it has enough information about the new workspace extended in its topology. Note that Master-DTSA is a server and maintain a lot of information about workspaces in its topology. These information can be stored by using common or distributed databases. Some procedures called in Figure 5 are not presented here and future works can implement the procedure using different techniques. However, two procedures are important to present: lookup and requestConfiguration, because both need specific messages on routing protocol.

Figure 6 shows lookup procedure and 7 shows requestCon-

 ISBN: 978-1-61208-398-8

```
1: procedure ONRECEIVEWORKSPACEATTACH(title)
2:     attached ← true
3:     wksInfo ← QUERYINLOCALDB(title)
4:     if wksInfo ≠ null then
5:         if CHECKREQUIREMENTS(wksInfo) then
6:             attached ←CALCULATEROUTE(wksInfo)
7:             attached ←UPDATEFLOWTABLES(wksInfo)
8:             attached ←INSERTINTOLOCALDB(wksInfo)
9:             attached ←NOTIFYMASTER(wksInfo)
10:            attached ←CALCULATEROUTE(wksInfo)
11:                ▷ wksInfo can be modified in each call
12:        else
13:            attached ← false
14:        end if
15:    else
16:        pathList ← LOOKUP(title)
17:        if pathList ≠ null then
18:            path ← CHECKBESTPATH(pathList)
19:            attached ← REQUESTCONFIGURATION(path)
20:            attached ←CALCULATEROUTE(wksInfo)
21:            attached ←UPDATEFLOWTABLES(wksInfo)
22:            attached ←INSERTINTOLOCALDB(wksInfo)
23:            attached ←NOTIFYMASTER(wksInfo)
24:            attached ←CALCULATEROUTE(wksInfo)
25:        else
26:            attached ← false
27:        end if
28:    end if
29:    return attached
30: end procedure
```

Figure 5. Routing Algorithm Procedure

figuration functions. Both are executed in Master-DTSA.

In the procedure shown in Figure 6, the checkTopology is invoked for the Master-DTSA verifies if its own topology supports the requirements specified by the Entity. The call for lookup is necessary if the Master does not contain informations about the required workspace. The addOwnPath routine is responsible to verify if the DTSA path is related only to that Master.

```
    procedure ONWORKSPACELOOKUP(title)
2:      pathList ← QUERYINLOCALDB(title)
        if pathList ≠ null then
4:          finalList ← pathList
        else
6:          supported ← false
            supported ← CHECKTOPOLOGY()
8:          if supported then
                pathList ← LOOKUP(title)
10:             if pathList ≠ null then
                    finalList ← ADDOWNPATH(pathList)
12:             end if
            end if
14:     end if
        return finalList
16: end procedure
```

Figure 6. WORKSPACE_LOOKUP Procedure

The procedure requestConfiguration shown in Figure 7 can run in a DTSA or in a Master DTSA. When it runs in Master-DTSA, this will notify all DTSAs in its topology that they are

```
    procedure REQUESTCONFIGURATION(wks, entity)
        routeInformation ← DEFINEROUTE()
3:      UPDATEFLOWTABLES(routeInformation)
        INSERTINTOLOCALDB(routeInformation, wks)
        NOTIFYMASTER(wks)
6:  end procedure
```

Figure 7. WORKSPACE_CONFIGURATION Procedure

on the chosen path, and notify the public control workspace to other Master DTSAs notify their DTSAs that they are on the path. The requestConfiguration can also arrive in a DTSA, and in this case the procedure shows the behavior as shown in the Figure 7.

Note that, in the requestConfiguration procedure, presented in Figure 7, the wks and entity are objects containing informations respectively about the workspace and the Entity that required the attach. The function defineRoute is responsible to define the best path considering all of the NE controlled by the DTSA. With this information, it is possible run updateFlowTables, a function where the DTSA sends to each NE in the route rule information to be inserted in flow tables of the NEs.

Considering the procedures, and the functions in each procedure, Table I presents the new primitives that are related with routing at ETArch.

TABLE I. ETArch ROUTING RELATED PRIMITIVES.

Message	*Description*
WORKSPACE_LOOKUP	Used when a DTSA does not have information about a workspace sought.
DTS_NOTIFY	Send from DTSA to Master. It is responsible to inform that a new workspace was created or extended by a DTSA in that Master topology.
WORKSPACE_CONFIGURATION	A DTSA sends this message to the private control workspace informing new extension of a workspace. Each DTSA that receives this message should configure the NE in its own topology necessary for this extension.

The procedures of routing protocol send the messages specified in Table I to the public or private control workspace. When a message, with one of these headers, arrives to a NE, it is forwarded by the control workspaces to a DTSA or Master-DTSA.

With the procedures and messages described in this section, it is possible to ETArch makes communication with two or more DTSAs, i.e., several controllers, defining the route on control plane, before the data starts to be forwarded in data plane. It is important to mention that the parameter attached found in Figure 5 means: the workspace is extended until the requesting Entity, i.e., the Entity is now attached with the workspace.

All procedures presented are subject to a rollback whether one of the functions fails.

V. CONCLUDING REMARKS AND FUTURE WORK

The control plane of the ETArch architecture is provided by the DTS, which is implemented through one or more agents named DTSAs. Each DTSA is responsible by a set of network elements whose topology is driven by local communication needs. In this work, by regarding routing, ETArch procedures and messages were described in order to provide communications using different DTSAs. It is important to remark that, up

 ISBN: 978-1-61208-398-8

to this moment, all the communications were made using only one DTSA, i.e., there was no structure to support a workspace over two or more DTSAs.

This paper presented the specification and design of the ETArch routing control plane procedures e rules. The proposed protocol will be incorporate with ETArch modules already deployed. For future works, the requirements as QoS, bandwidth, secure and energy consumption will be incorporated to provide scenarios where it will be possible to take tests and experiments to run the specified algorithms presented here.

Despite of the current work be made for ETArch, it can be applied to the research in SDN as a possible scenario for the routing by using the control plane, by improving the processing time spent in the routing over the current Internet. For the ETArch, this work provides the capability of using different DTSAs (controllers) in the workspace communications. This is an important feature, because, for example, it is not interesting for a carrier providing information about its topology and network elements.

The work presented here follows the SDN concepts, separating the control and data plane, and the routing proposed here is able to define the complete path before the data starts being forwarding. The routing time is spent during the path establishment and the communications phase spend only switching time, differently of today's Internet.

ACKNOWLEDGMENT

This work has been partially funded by the Brazilian agencies: CAPES; CNPq; FAPEMIG; and PROPP/UFU - and by ALGAR Telecom.

REFERENCES

[1] V. G. Cerf and E. Cain, "The DoD internet architecture model," Computer Networks (1976), vol. 7, no. 5, Oct. 1983, pp. 307–318.

[2] T. Zahariadis et al., "Towards a future internet architecture," in The Future Internet. Future Internet Assembly 2011: Achievements and Technological Promises, ser. Lecture Notes in Computer Science, J. Domingue, A. Galis, A. Gavras, T. Zahariadis, and D. Lambert, Eds. Berlin, Heidelberg: Springer-Verlag, 2011, vol. 6656, pp. 7–18.

[3] N. S. Foundation. NSF future internet architecture project. [Online]. Available: http://www.nets-fia.net/ [retrieved: Mar., 2015]

[4] Eurescom. Future internet assembly - european future internet portal. [Online]. Available: http://www.future-internet.eu/home/future-internet-assembly.html [retrieved: Mar., 2015]

[5] J. de Souza Pereira, F. de Oliveira Silva, E. Filho, S. Kofuji, and P. Rosa, "Title model ontology for future internet networks," in The Future Internet, ser. Lecture Notes in Computer Science, J. Domingue et al., Eds. Springer Berlin Heidelberg, 2011, vol. 6656, pp. 103–114.

[6] F. Silva et al., "Entity title architecture extensions towards advanced quality-oriented mobility control capabilities," in Computers and Communication (ISCC), 2014 IEEE Symposium on, June 2014, pp. 1–6.

[7] M. A. Goncalves, P. F. Rosa, and de Oliveira Silva., "Multicast traffic aggregation through entity title model," in The Tenth Advanced International Conference on Telecommunications (AICT), ThinkMind, Ed. IARIA, 2014, pp. 170–180.

[8] C. Guimaraes, D. Corujo, F. Silva, P. Frosi, A. Neto, and R. Aguiar, "Ieee 802.21-enabled entity title architecture for handover optimization," in Wireless Communications and Networking Conference (WCNC), 2014 IEEE, April 2014, pp. 2671–2676.

[9] W. Fokkink, Distributed Algorithms: An Intuitive Approach. Cambridge, Massachusetts: The MIT Press, Dec. 2013.

[10] D. J. W. A. S. Tanenbaum, Computer Networks, international ed of 5th revised ed edition ed. Harlow, Essex: Pearson Education Limited, Jul. 2013.

[11] D. Massey, L. Wang, B. Zhang, and L. Zhang, "A scalable routing system design for future internet," in Proc. of ACM SIGCOMM Workshop on IPv6, 2007.

[12] K. Calvert, J. Griffioen, and L. Poutievski, "Separating routing and forwarding: A clean-slate network layer design," in Fourth International Conference on Broadband Communications, Networks and Systems, 2007. BROADNETS 2007, Sep. 2007, pp. 261–270.

[13] F. Le, G. G. Xie, and H. Zhang, "On route aggregation," in CoNEXT, K. Cho and M. Crovella, Eds. ACM, 2011, p. 6.

[14] S. Strowes, Compact Routing for the Future Internet. University of Glasgow, 2012.

[15] X. Yang, D. Clark, and A. W. Berger, "Nira: A new inter-domain routing architecture," IEEE/ACM TRANSACTIONS ON NETWORKING, 2007.

[16] I. A. Ganichev, "Interdomain multipath routing," Ph.D. dissertation, EECS Department, University of California, Berkeley, Dec 2011.

[17] G. T. Nguyen, R. Agarwal, J. Liu, M. Caesar, P. B. Godfrey, and S. Shenker, "Slick packets," SIGMETRICS Perform. Eval. Rev., vol. 39, no. 1, Jun. 2011, pp. 205–216.

[18] B. Parno, M. Luk, E. Gaustad, and A. Perrig, "Secure sensor network routing: A clean-slate approach," in Conference on Future Networking Technologies (CoNEXT), December 2006.

[19] K. L. Calvert, J. Griffioen, and L. Poutievski, "Separating Routing and Forwarding: A Clean-Slate Network Layer Design," in In proceedings of the Broadnets 2007 Conference, September 2007.

[20] Y. Wang, F. Esposito, I. Matta, and J. Day, "Recursive InterNetworking Architecture (RINA) Boston University Prototype Programming Manual (version 1.0)," CS Department, Boston University, Tech. Rep. BUCS-TR-2013-013, November 11 2013.

[21] F. de Oliveira Silva, M. Goncalves, J. de Souza Pereira, R. Pasquini, P. Rosa, and S. Kofuji, "On the analysis of multicast traffic over the entity title architecture," in 2012 18th IEEE International Conference on Networks (ICON), 2012, pp. 30–35.

[22] ON.LAB. ONOS - open network operating system. [Online]. Available: http://tools.onlab.us/onos.html [retrieved: May, 2014]

[23] ——. ONOS at ONS 2014. [Online]. Available: http://www.slideshare.net/ON_LAB/onos-at-ons-2014 [retrieved: Mar., 2014]

[24] I. Seskar, K. Nagaraja, S. Nelson, and D. Raychaudhuri, "MobilityFirst future internet architecture project," in Proceedings of the 7th Asian Internet Engineering Conference, ser. AINTEC '11. New York, NY, USA: ACM, 2011, p. 13.

[25] S. C. Nelson, G. Bhanage, and D. Raychaudhuri, "GSTAR: generalized Storage-Aware routing for mobilityfirst in the future mobile internet," in Proceedings of the sixth international workshop on MobiArch, ser. MobiArch '11. New York, NY, USA: ACM, 2011, p. 1924.

[26] J. C. Lema et al., "Evolving future internet clean-slate entity title architecture with quality-oriented control plane extensions," in The Tenth Advanced International Conference on Telecommunications (AICT), ThinkMind, Ed. IARIA, 2014, pp. 161–167.

 ISBN: 978-1-61208-398-8

Toll Fraud Detection in Voice over IP Networks Using Communication Behavior Patterns on Unlabeled Data

Sandra Kübler, Michael Massoth, Anton Wiens and Torsten Wiens
Department of Computer Science
Hochschule Darmstadt – University of Applied Science
Darmstadt, Germany
e-mail: {sandra.kuebler | michael.massoth | anton.wiens | torsten.wiens}@h-da.de

***Abstract*—Widespread monetary losses are known to be caused worldwide by fraud attacks on Voice over IP systems. In 2014, several millions of FRITZ!Box routers have been compromised and used to conduct phone calls to international destinations. By using fraud detection systems, such attacks can be detected. By analyzing Call Detail Records (CDRs), various algorithms can be applied to detect fraud. Unfortunately, this data is mostly unlabeled, meaning no indications on which calls are fraudulent or non-fraudulent exist. In this work, a new method to detect fraud is presented, utilizing the concept of clustering algorithms leading to *behavior pattern recognition* using information retrieved from user profiles. The grouping aspect of clustering algorithms regarding the similarity of objects leads to data depicting the behavior of a user to be matched against behavior patterns. If a deviation from the assigned behavior patterns occurs, the call is considered fraudulent. A prototype has been implemented with two behavior patterns defined, making it possible to detect fraud. It can further be refined by adjusting multiple thresholds, as well as defining more behavior patterns. The prototype is to be integrated in an existing fraud detection system of Hochschule Darmstadt, being developed in cooperation with a small and medium-sized enterprise (SME) telecommunication provider, improving the quality of its VoIP services.**

Keywords-Fraud detection; Voice over IP networks; behavior pattern recognition; unlabeled data; FRITZ!Box.

I. INTRODUCTION

Voice over IP (VoIP) has been well-established as one of the possibilities to perform voice communication. As it uses the internet as a means of data transportation, it also inherits its drawbacks, also concerning its security flaws. These security flaws can be exploited by criminals by, for instance, taking over a private branch exchange (PBX) and performing fraudulent phone calls using a specific user's account. Telecommunication providers, especially small and medium-sized enterprises (SME), suffer from those attacks as they lead to financial losses and a decrease of trust on part of their customers.

Every two years, the Communications Fraud Control Association (CFCA) conducts a survey on global fraud loss. The 2013 survey shows that approximately 46.3 billion USD have been lost due to fraud attacks, denoting an increase of 15% in comparison to 2011 [1].

The actuality of fraud attacks in VoIP is further emphasized by the "FRITZ!Box incident". *AVM FRITZ!Boxes* are multifunctional routing devices, which are very popular in Germany. In February 2014, several million units have been compromised by hackers exploiting security vulnerabilities [2]. For instance, this caused a regional German telecommunication provider financial losses of more than 200,000 € during one month [3].

Meanwhile, the security vulnerabilities have been patched by the manufacturer, but users still can be affected, as it is very likely that sensitive data (e.g., login data) has been stolen as well. If the password has not been changed, the system may still be vulnerable. Furthermore, the update requires manual patching. This is further accentuated in [4], where it is shown that users who did not patch their units are still suffering from attacks.

Fraudulent activities in the telecommunication sector can be countered using various mechanisms. Possibilities range from techniques based on user profiling where deviations from a user's normal behavior are considered fraudulent, using machine learning algorithms or even combining techniques from various fields and developing frameworks with additional features as a means to prevent fraud in the first place [5]-[9].

The *University of Applied Sciences Darmstadt* aims to detect and therefore minimize financial loss with its research project "Trusted Telephony". Furthermore, it intends to provide enhanced security in VoIP telecommunication, leading to a versatile *fraud detection system*, which is currently in development. It utilizes various techniques gained from ongoing research on fraud detection, e.g., using rule-based and user profiling techniques. The work at hand is part of the research project "Trusted Telephony" and is preceded by the works [6][10][11]. The German telecommunication service provider *toplink GmbH* cooperates with the *University of Applied Sciences Darmstadt*, especially providing the necessary data for analysis.

In this paper, a new method to detect fraud in VoIP communication is presented. This new method is intended to be a new component for the fraud detection system. It is based on the findings on fraud detection obtained from preceding work [6], which dealt with the issues of the FRITZ!Box incident as well. A summary of the most important findings concerning an analysis on data obtained during the FRITZ!Box incident is given in Section V.

The idea of the new approach is to adapt the idea of the concept of *clustering* ("grouping" of data based on the

 ISBN: 978-1-61208-398-8

similarity of an object) from machine learning and combine it with *user profiles*, leading to an approach based on *behavior pattern recognition* using pieces of information retrieved from user profiles. The thought of potentially using clustering algorithms in the first place arose because no labeled data was available. Therefore, techniques not solely relying on the existence of labeled data became more interesting to the project.

A. Call detail records

In this work, the data provided by *toplink GmbH* is in the form of Call Detail Records (CDRs). These text files contain call parameters, e.g., caller- and callee-party parameters, starting time and call duration.

B. Structure of the paper

The introduction is followed by an overview of related work in Section II. Section III gives a brief overview of unsupervised learning, as it is relevant for the concept to detect fraud cases. The basic idea of user profiling is described in Section IV. Section V describes the data collected during the FRITZ!Box incident, followed by a use case of the presented method in Section VI. The concept of *communication behavior patterns* using data from user profiles is described in detail in Section VII. The prototypical implementation is described in Section VIII, including information about the utilized data set, the experimental setup and its results. A conclusion to this paper is presented in Section IX, being followed by possible future work in Section X.

II. Related Work

As a means to visualize user accounts, self-organizing maps (SOM) are used in [5]. This visualization is used to differentiate between normal and fraudulent ones. Three features are extracted from the CDR data and used for analysis: Call destination, call start time and call duration. According to the authors, the method has a true positive rate (TPR) of 90% and a false positive rate (FPR) of 10%.

In order to cluster probabilistic models, a framework for self-organizing maps has been developed by Hollmén, Tresp and Simula [12]. User profiles using data of mobile communication networks have been used for test runs of the system. The output is presented visually, so that the fraudulent calls can be distinguished from normal ones.

The authors of [7] focus on the detection of superimposed fraud using two signature methods, each summarizing a user's behavior. The first presented approach is based on a deviation of the user's current behavior and his signature, while the second is based on a dynamic clustering analysis. In the second approach, a sudden change or "shift" of a user's signature from one cluster to another is the criterion for a classification as fraud. The similarity between a signature and a cluster centroid, which in itself is defined as a signature, is crucial for such a shift. The detection rates of both methods have been estimated: The first one promises a TPR of 75% and the second one a TPR of 91%. Also, a combination of both approaches is examined.

The framework *SUNsHINE*, which is able to detect and prevent VoIP fraud by combining real-time capable components with an offline statistical analysis, is presented in [9]. Multiple data sources, network traffic data and CDRs, can be used. Different algorithms and techniques are used, e.g., rule sets, profiling, neural networks and clustering. No estimations concerning the detection rate are given.

As some of the related work is using neural networks or variations of these, it should be further pointed out that one of the major drawbacks of using neural networks lies in the necessity of having labeled data for training. While it is possible to use SOMs in the sense of clustering, these still require some kind of "training" or evaluation.

The preceding works [11] and [6] as part of the research project had to deal with this problem as well. In [11], a detailed list regarding related work based on user profiling is provided and a method based on statistical user profiling is presented. Two user profiles containing statistical features are generated, representing the past (Past Behavior Profile, PBP) and the present (Current Behavior Profile, CBP), using a significant deviation of a user's behavior in contrast to his past behavior as an indication for possible fraud. The idea of using two user profiles is based on [8] and [13], which both use a user profile history and a current user profile. In [11], a TPR of 90% and a FPR of 1.22% is estimated.

The successor of the approach described above is using an enhanced approach in order to deal with the fraud cases acquired during the FRITZ!Box incident [6]. Distributed fraud attacks, as described therein, can be detected by profiling the destination numbers instead of a user as it is normally done when using the principles of user profiling. The approach described in [6] differs in the *point of view* of the data in contrast to this work. The work at hand has been inspired by the concept of clustering algorithms, as the aspect of *finding similarities* has been adopted.

III. Unsupervised Learning

One conceptual requirement for the component being developed for this work is to be able to detect fraud without using labeled data. Hence, algorithms based on unsupervised learning immediately suggest themselves [14][15]. Clustering techniques, grouping similar objects, are most commonly used. The similarity function used depends on the type of clustering algorithm, e.g., hierarchical or centroid based methods. Further information on clustering algorithms can be found in [15].

In this work, as a means to perform unsupervised learning, user profiling is being applied (see Section IV) for fraud detection. Applying user profiles proved successful in previous work on the fraud detection system [6][11], as the data is put into a user context which is missing otherwise. This context is important as not every user behaves the same.

The new component should be integrated into an existing framework, which should function in nearly real-time. The authors decided to use clustering algorithms as a "preprocessing step" during the data analysis, as opposed to related work, e.g., [5]. This is due to the fact that a lot of data has to be processed and the clustering algorithms have to be evaluated. Additionally, clustering algorithms can also be

 ISBN: 978-1-61208-398-8

time-consuming. The pieces of information obtained through clustering are used to obtain indications of the definition of behavior patterns and thresholds. The definition of behavior patterns is a key part in the work at hand, as each pattern describes a distinct behavior of a user and as the matching to a behavior pattern and its growth are used for the actual fraud detection. For the preprocessing step, clustering algorithms (k-means, unsupervised SOM) as they are implemented in the tool WEKA, which provides machine learning algorithms for data mining tasks [16], are applied. Furthermore, ideas derived from clustering techniques in general influenced the actual concept of *communication behavior patterns* using data from user profiles as described in Section V.

IV. USER PROFILING

Two types of analysis exist: absolute and differential [17]. While an absolute analysis is retrieving pieces of information directly from CDRs and therefore is in need of having a firm understanding of fraud patterns, a differential analysis summarizes the retrieved information into statistical features over a distinct period of time. The latter is also called *behavior-* or *user profiling*. Utilizing this profile, it is possible to identify a change in a user's behavior over a given period of time. The utilization of user profiling, varying in its concept and features used, is addressed in related work [6][8][11][18][19], which partially use the common features *duration per call*, *number of calls per customer* and *costs per call*.

V. FRITZ!BOX INCIDENT DATA

Fraud cases originating from the FRITZ!Box incident share some common characteristics. These characteristics have been described in detail in [6] and will be briefly summarized, as well as complemented in the following. A description of how these units could have been taken over is given in [6]. Occurring attack patterns are as follows:

- Different and numerous international numbers have been dialed in rapid succession which were either not connected (call attempts) or having a short duration (call connects).
- From the user's perspective, only one international number has been dialed either resulting in a call attempt or a call connect. From the call destinations' perspective, up to eight different users dialed the number. The numbers have been dialed in rapid succession (at night-time in one-second intervals), having a mean duration of approximately 7-8 minutes.
- A user dials several, mostly international numbers. A destination number is being dialed by 3 users on average.
- While national numbers have been dialed, call attempts, as well as call connects to international numbers were made nearly simultaneously.
- Most fraudulent calls were made between afternoon and early morning, showing a peak in the night-time.

Especially countries from zones 2 (mostly Africa) and 3 (mostly Europe) have been called numerously. The listed criteria can occur combined. The amount of call attempts outweighs the amount of call connects. The duration of the phone calls ranges from approximately 30 ms up to 11 minutes.

VI. USE CASE – COUNTRY PROFILING

In the following, a use case for the concept of *communication behavior patterns* using information from user profiles is depicted (see Fig. 1).

A customer of a telecommunication provider – in most cases in the given data set, a customer equals a company – conducts business calls to various foreign countries. These international calls are further described as matches to behavior patterns of the customer, each being a differentiation of a behavior pattern describing international calls in general. In Fig. 1, the size of each ellipse surrounding a country A to E indicates how many calls are usually – i.e., as it had been profiled during the initialization phase – conducted to the destination. Now, a new behavior pattern reflecting the behavior to conduct calls to country E emerges. During a short time span, e.g., of one hour, the number of matches to this new behavior pattern grows, indicating that distinctly more calls have been conducted to country E. Furthermore, this new behavior pattern and the growth of the match to it over a short period of time could indicate fraudulent calls, as typically, the customer does not conduct that many calls to country E so that a match to such a behavior pattern could be justified.

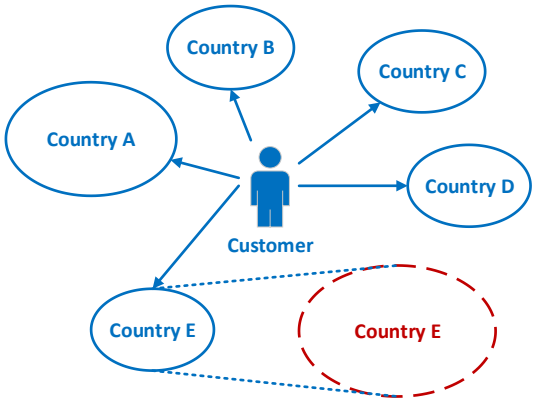


Figure 1. Depiction of the use case with a customer having his international groups and a new one growing over a short time interval, indicating fraud.

This use case illustrates the *potential* of the concept of *communication behavior pattern recognition* with information from user profiling, as it is described in detail in the following section, and how it can be used for fraud detection. Instead of profiling by country, it would also be possible to profile by telecommunication providers, as our data suggest.

VII. CONCEPT OF COMMUNICATION BEHAVIOR PATTERNS

The idea behind the concept of using *behavior patterns* with information from user profiles is to adapt the principle

 ISBN: 978-1-61208-398-8

of *clustering* algorithms, combined with the usage of *user profiles* (differential analysis). Thus, the requirement of the concept to function with unlabeled data can be met. The concept itself can therefore be categorized as an unsupervised classification method.

Similar objects following similar patterns are to be assigned into the same behavior pattern. To associate with a behavior pattern, each shall have its own criteria, where similar groups possess similar criteria. In this work, user profiles, using the information retrieved from the CDRs as a base for their features, are used as objects. In order to describe the behavior concerning a distinct aspect of a user or a group of users, the calls of a user profile are matched against predefined behavior patterns. A user is able to have matches to several behavior patterns.

To obtain an indication for the thresholds and to search for behavior patterns, clustering algorithms from WEKA were used.

A. Data preparation

Only a fraction of the information contained in the CDRs is used as input for a user profile, as not all information concerning a VoIP connection is necessary for analysis, as shown in [11]. The important pieces of information extracted from a CDR are the following attributes A_1 to A_4:

A_1 User ID
A_2 Timestamp of the call
A_3 Duration of the call
A_4 Destination number

The session ID is not used for the construction of a user profile, but as a unique identifier of the corresponding CDR. The information obtained from A_4 is further categorized into its call region *national*, *mobile* and *international*. The information whether the call had been connected or was merely a call attempt is being retrieved from A_3.

A_2 is further processed and divided into more fine-grained pieces of information, namely whether or not the call occurred on a weekend and if the call has been made during work hours (7:00 am to 18:59 pm) or after hours (19:00 pm to 6:59 am) with a time span of 12 hours each. This segmentation is done because of findings from the analysis described in Section V, as a considerable amount of fraudulent calls, especially call attempts, has been conducted at night. Furthermore, this allows for a more versatile definition and use of behavior patterns without being too complex.

B. User profiles

A user profile contains data extracted from CDRs (see above) related to a user over a certain period of time *t*.

After being filled with data accumulated during *t*, a *user profile* is considered ready to test for fraud. For *t*, at least one week is considered appropriate [6][7][8][9][11], as substantial data about the normal behavior of a user has to be gathered, resulting in a "training phase" of a user profile.

CDRs outdating *t* are removed from a profile. Based on the data contained in a profile, features can be extracted.

C. Behavior patterns

As mentioned before, a *behavior pattern* reflects a distinct behavior or rather a behavior in a specific context of a user. For instance, if a user is calling international destinations often, this user matches the behavior pattern "International Calls". Another specific context is that calls have been made on weekend or during work hours. It is possible for a user to *match* one or more behavior patterns.

1) Features

A behavior pattern has its own defining set of features *F* called *feature vector*, with comparable behavior patterns having similar defining features. As these features are highly dependable on the context or rather the criteria of a behavior pattern, an overall definition for a feature vector cannot be given. The features are derived from the data contained in a user profile. Essentially, there are two types of features: numeric and Boolean (true/false).

Examples for two behavior patterns, their criteria and therefore feature vectors:

- *"International Calls After Hours"*: The criteria for this pattern are: The call has to be connected, the call region is *international* and the call is made *after hours*.
- *"Weekend Calls"*: The only criterion is for the calls to be made on a weekend.

For both behavior patterns applies that the single numerical value in the feature vector is the accumulation of the respective calls during a time span t_{BP}. For t_{BP}, a value of one hour has been chosen, as this time span is neither too short nor too long.

2) Criteria for a behavior pattern match

In order for a *user* to *match* a behavior pattern, every feature of a feature vector, depending on its type, has to meet its criteria:

- Numeric: A statistical or numeric feature has to pass a *threshold*.
- Boolean: A Boolean feature has to have the value *true*.

For every defined behavior pattern, the criteria are tested. This way, it is possible for a CDR of a user to lead to a match to more than one behavior patterns.

3) Metric for a match

All calls matching a distinct behavior pattern are stored in respective lists. Over time, the length of such a list - and, therefore, the *grade* of a match - can diminish or grow. This is further denoted as a *growth* of a match to a behavior pattern.

The *growth G* of a match to a behavior pattern over a timespan is measured as:

$$G = \frac{C_L}{\bar{x}\,(C_P)} \quad (1)$$

C_L denotes a list of all connected calls during the current (latest) hour and C_P a list of all connected calls in the past.

 ISBN: 978-1-61208-398-8

For both C_L and C_P, calls from the list of matches are used. $\bar{x}$ denotes the arithmetic mean over the respective list.

4) Change of a match

The *growth G* of a match to a behavior pattern described above is further used as a criterion to mark a current call as *fraudulent*, as it is defined in the following case differentiation:

$$Fraud = \begin{cases} true, G > T_{BP} \\ false, otherwise \end{cases} \quad (2)$$

T_{BP} denotes a threshold for the growth of a match to a behavior pattern. If T_{BP} is passed, the current call, which had been causal for *passing* the threshold, is the first call to be considered fraudulent. All subsequent calls which are still triggering *true* are considered fraudulent as well. Additionally, a *weight* can be assigned to every behavior pattern, indicating how much a growth of a match influences the assignment of a call as fraudulent. This leads to an enhancement of the case differentiation (2):

$$Fraud = \begin{cases} true, G \cdot w > T_{BP} \\ false, otherwise \end{cases} \quad (3)$$

VIII. PROTOTYPE

The concept described in Section VII has been implemented as a prototype. The description of the prototype consists of the used data set (Subsection A), the experimental setup (Subsection B) and the results (Subsection C).

A. Used data

Real life traffic data over a time span of seven weeks provided by *toplink GmbH* has been used to test the prototypical implementation. For the initialization of the user profiles, as well as behavior patterns of a user, the data of the first week has been used, as they contained no known fraudulent activity. Out of the seven weeks, there is at least one week included with definite fraud attacks having the pattern described in Section V. The rest of the data shows partial signs of the FRITZ!Box fraud attack pattern as well.

The data set comprises 10,401,547 CDRs. As only outgoing calls, as well as successfully connected calls (call connects) are of importance, 2,749,860 CDRs were left.

B. Experimental setup

For the prototypical implementation, two simple behavior patterns have been defined:

- *IntCallsPattern*: All connected calls having an international destination match the behavior pattern.
- *IntCallsAfterHoursPattern*: All connected calls having an international destination and having been conducted in the after hours match this behavior pattern.

The thresholds for the statistical features for both behavior patterns, as well as indications about the thresholds concerning the change of a match to a behavior pattern have been derived using clustering algorithms from WEKA. The applied clustering algorithms were k-means, EM and an implementation of a SOM as a clustering algorithm.

C. Results

Determining a True and False Positive Rate (TPR and FPR) poses a difficult task if only unlabeled data is available. Due to the analysis performed on the data retrieved during the FRITZ!Box incident, an approximation concerning the TPR was possible. Nevertheless, not all fraudulent data has been known during the evaluation of the prototype. The data set described in Section VIII.A has been used. The following steps have been applied:

1. Apply the thresholds and weight values retrieved from clustering algorithms and given from experience, respectively.
2. Run the prototype with the defined two behavior patterns.
3. Analyze the results utilizing the knowledge derived from the analysis of the data, as well as from *toplink*.

In total, 17,110 fraud cases were reported and analyzed. During the analysis, one customer was noticeable in his behavior to conduct calls to foreign destinations very often, even not during the timeframe of the FRITZ!Box incident. Therefore and because of other aspects found in our analysis, this customer can be considered being a call center. Such a call center is a likely candidate to be added to a whitelist and thus can be ignored, leading to a total of 13,503 reported fraud cases if subtracted. The TPR measured is *98.4%*. The TPR would vary if this specific customer would be taken into account, but this type of customer could easily be excluded in preprocessing via whitelisting. The measured FPR is below 0.01 %.

Surely not all fraud instances of the FRITZ!Box incident could be found. This can be said even though not enough labeled data existed, as valuable time – and therefore, CDRs - passes in order for a user to match a behavior pattern and be associated with the described behavior. Afterwards, a threshold concerning the growth of a match has to be passed, resulting in an equivalent to a "settling-in phase". Thus, it is possible that not all fraudulent instances were detected.

TABLE I. FPR AND TPR COMPARISON WITH RELATED WORK

	TPR	**FPR**
This work	98.4%	< 0.01%
Previous work [6]	95% (100%)	0.7%
Previous work [11]	90%	1.22%
Related work [5]	90%	10%

Table I. shows a comparison of the TPR and FPR with related work. Concerning the results in [6], the TPR had been reduced from 100% to 95% by the authors, based on a qualified estimation, as it is possible that not all fraudulent calls in the dataset are known. Concerning the work at hand and the dataset used, this could be very likely, too.

 ISBN: 978-1-61208-398-8

IX. CONCLUSION

It is possible to detect fraud attacks using the presented approach. Regarding the FRITZ!Box data, it is comparable in quality to the approach presented in [6]. With the possibility to define behavior patterns, it has the additional potential to be more versatile. Currently, there are only two behavior patterns defined, but with the addition of more behavior patterns and better regulated thresholds, the results can be improved as more fraud patterns could be detected. Also, more heterogeneous patterns could be found.

X. FUTURE WORK

One possibility to improve the detection rate is to increase the time span for the initialization phase. Behavior patterns relating to weekends or workdays are only meaningful for a stable analysis if there are at least three weeks of data given.

The importance of whitelisting is shortly mentioned in Section VIII.C, as a customer being most likely a call center causes the TPR to differ significantly.

Furthermore, the idea of including a "global trend" arose, similar to the one presented in [11] where a global profile possessing the CDRs of all users has been included in order to balance fluctuations in the data. In the work at hand, the growth of lists of calls matching distinct behavior patterns can be monitored globally. Concerning the users, two possibilities have to be considered:

- The user only just did not meet the requirements of having enough matches to a behavior pattern to be associated with a pattern and
- The user has enough matches to a behavior pattern to be associated with it, but only just did not pass the threshold concerning the growth and therefore the call is not considered fraudulent.

If such a distinct behavior pattern shows fraudulent activity originating from several users, this "global trend" can influence the result concerning the aforementioned users.

Including information given by call attempts and call termination cause codes can further improve the detection result. They can provide insight whether a fraudulent attack is currently prepared or conducted. Additionally, "normal" behavior patterns - e.g., "National Calls" - have to be considered as well. Their sole existence can be used as a further criterion for other behavior patterns.

Furthermore, the possibility of conflicting behavior patterns can be considered as well. For instance, a user usually calls on weekends and this behavior had been learned during the initialization phase. Suddenly – e.g., during one hour – a new matching of a behavior pattern reflecting workday calls emerges. In this case, the existence of the workday behavior pattern matching conflicts with the weekend pattern matching and can be considered suspicious behavior.

Regarding the machine learning part of this concept, further clustering algorithms are currently being evaluated to improve the process of retrieving the thresholds, as well as the values themselves. Furthermore, techniques like *Principal Component Analysis* (PCA) ought to be used for first-step data analysis, as preliminary tests on raw CDR data suggest.

ACKNOWLEDGMENT

We want to express our gratitude to the State of Hesse, Germany and its research program LOEWE for providing funding. Additionally, we want to thank *toplink GmbH* (Darmstadt, Germany) for their cooperation and for providing the data for our research project.

REFERENCES

[1] Communications Fraud Control Association, "Global Fraud Loss Survey," October 2013. [Online]. Available from: http://www.cfca.org/pdf/survey/CFCA2013GlobalFraudLossSurvey-pressrelease.pdf 2014.11.17.

[2] R. Eikenberg, "Hack on AVM routers: Fritzbox breach disclosed, millions of routers at risk," 07 03 2014. [Online]. Available from: http://www.heise.de/security/meldung/Hack-gegen-AVM-Router-Fritzbox-Luecke-offengelegt-Millionen-Router-in-Gefahr-2136784.html 2014.11.19.

[3] R. Eikenberg, "Change VoIP password now: criminals exploit captured Fritzbox data," 26 03 2014. [Online]. Available from: http://www.heise.de/security/meldung/Jetzt-VoIP-Passwort-aendern-Kriminelle-nutzen-erbeutete-Fritzbox-Daten-aus-2155168.html 2014.12.08.

[4] WAZ (DerWesten), "AVM warns against attacks on its Fritzbox routers," 29 09 2014. [Online] Available from: http://www.derwesten.de/ratgeber/avm-warnt-vor-angriffen-auf-seine-fritzbox-router-id9881008.html 2014.11.26.

[5] D. Olszewski, J. Kacprzyk, and S. Zadrożny, "Employing Self-Organizing Map for fraud detection" The 12th International Conference on Artificial Intelligence and Soft Computing (ICAISC 2013), June 2013, pp. 150-161, ISBN: 978-3-642-38657-2.

[6] A. Wiens, T. Wiens, and M. Massoth, "Approach on fraud detection in Voice over IP networks using call destination profiling based on an analysis of recent Attacks on Fritz!Box units" The Sixth International Conference on Emerging Network Intelligence (EMERGING 2014) IARIA, Aug. 2014, pp. 29-34, ISSN: 2326-9383, ISBN: 978-1-61208-357-5.

[7] R. Alves et al., "Discovering telecom fraud situations through mining anomalous behavior patterns". Proceedings of the DMBA Workshop on the 12th ACM SIGKDD, 2006.

[8] C. S. Hilas and P. A. Mastorocostas, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," Knowledge-Based Systems, vol. 21, issue 7 , pp. 721-726, Oct. 2008, doi:10.1016/j.knosys.2008.03.026.

[9] D. Hoffstadt et al., "A comprehensive framework for detecting and preventing VoIP fraud and misuse," 2014 International Conference on Computing, Networking and Communications (ICNC), Feb. 2014, pp. 807-813, doi:10.1109/ICCNC.2014.6785441.

[10] S. Augustin et al., „Telephony fraud detection in Next Generation Networks" The Eighth Advanced International Conference on Telecommunications (AICT 2012) IARIA, May 2012, pp. 203-207, ISSN: 2308-4030, ISBN: 978-1-61208-199-1.

[11] A. Wiens, T. Wiens, and M. Massoth, "A new unsupervised user profiling approach for detecting toll fraud in VoIP networks" The Tenth Advanced International Conference on Telecommunications (AICT 2014) IARIA, July 2014, pp. 63-69, ISSN: 2308-4030, ISBN: 978-1-61208-360-5.

[12] J. Hollmén, V. Tresp, and O. Simula, "A Self-Organizing Map for clustering probabilistic models" Ninth International Conference on Artificial Neural Networks (ICANN) vol. 2,

1999, pp. 946-951, ISSN: 0537-9989, ISBN: 0-85296-721-7, doi:10.1049/cp:19991234.

[13] P. Burge and J. Shawe-Taylor, "An unsupervised neural network approach to profiling the behavior of mobile phone users for use in fraud detection" Journal of Parallel and Distributed Computing 61, 2001, pp. 915-925, doi:10.1006/jpdc.2000.1720.

[14] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey" ACM Computing Surveys (CSUR), vol. 41, issue 3, pp. 15:1-15:58, Jul. 2009, doi:10.1145/1541880.1541882.

[15] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey" A Review of Machine Learning Techniques for Processing Multimedia Content, Report of the MUSCLE European Network of Excellence (6th Framework Programm), 2005.

[16] WEKA, Machine Learning Group at the University of Waikato, official homepage. [Online] Available from: http://www.cs.waikato.ac.nz/ml/weka/ 2014.12.15.

[17] P. Burge et al., "Fraud detection and management in mobile telecommunications networks" European Conference on Security and Detection (ECOS97) Incorporating the One Day Symposium on Technology Used for Combatting Fraud, 1997, pp. 91-96, doi:10.1049/cp:19970429.

[18] Y. Moreau, H. Verrelst, and J. Vandewalle, "Detection of mobile phone fraud using supervised neural networks: a first prototype" 7th International Conference on Artificial Neural Networks (ICANN), 1997, pp. 1065-1070, ISSN: 0302-9743, ISBN: 978-3-540-63631-1, doi:10.1007/BFb0020294.

[19] M. Taniguchi, M. Haft, J. Hollmén, and V. Tresp, "Fraud detection in communication networks using neural and probabilistic methods" Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998, vol. 2, pp. 1241-1244, ISSN: 1520-6149, ISBN: 0-7803-4428-6, doi:10.1109/ICASSP.1998.675496.

 ISBN: 978-1-61208-398-8

ACROSS-FI: Attribute-Based Access Control with Distributed Policies for Future Internet Testbeds

Edelberto Franco Silva
Institute of Computing
Fluminense Federal University (UFF)
MídiaCom Laboratory
Niterói, RJ, Brazil
Email: edelberto@midiacom.uff.br

Natalia Fernandes Castro
Dept. of Telecommunications Engineering
Fluminense Federal University (UFF)
MídiaCom Laboratory
Niterói, RJ, Brazil
Email: natalia@midiacom.uff.br

Débora C. Muchaluat Saade
Institute of Computing
Fluminense Federal University (UFF)
MídiaCom Laboratory
Niterói, RJ, Brazil
Email: debora@midiacom.uff.br

***Abstract*—Interests in access control authorization methods for distributed resources have been growing as more shared resources environments and resource federations have been made available, both in academy and in industry. Different proposals aiming at creating a granular and scalable access control in those distributed environments have been presented in the literature. The standardization of access control models based on roles and attributes are examples of that effort. However, none of the existing proposals or standards present a complete authentication and authorization framework that can be adapted for different distributed environments. This work presents an authentication and authorization framework based on policies and attribute aggregation for controlling access into Future Internet (FI) distributed testbeds. A generic solution for attribute-based access control in Future Internet testbeds federation is implemented and evaluated, providing a generic interface to allow communication between the FI resource federation and our access control proposal. Based on user and resource's attributes, policies are dynamically applied to control which resources a user may require. This work has been validated in an experimental identity management laboratory (GIdLab) enabling the use of identity management services offered in an academic identity federation and in an experimental environment for the Future Internet.**

Keywords*—future internet; authorization; authentication; attribute-based access control.

I. INTRODUCTION

Identity Management (IdM) is the set of processes and technologies used for guaranteeing the identity of an entity. IdM ensures the quality of identity information such as identifiers, credentials, and attributes and uses it for authentication, authorization, and accounting processes [1].

Authentication procedures focus on confirming the identity of an entity, that is, checking that an entity is who it claims to be. Authorization mechanisms define the access rights to resources associated to an identity. Authorization procedures describe these access rights to ensure that they are obeyed. Finally, accounting refers to track network resource consumption by users for capacity planning and billing.

In recent years, the use of academic authentication and authorization (A&A) federations to control access to resources became popular [2][3]. In Brazil, for instance, academic researchers access scientific publication repositories using identities of the CAFe academic federation [4]. Hence, there is no need to duplicate user information in local databases.

The Federated Identity Management (FIM) is the basis of this work, when users from many institutions can access services provided by other partner institutions. A federated and distributed identity service depends on the ability of any service provider to trust the credentials provided to them by other entities. In this scenario, IdM appears as a strong requirement for establishing the trust environment among participants, as to share tools or resources among each other.

Important examples of such environments are the initiatives for experimental facilities for the Future Internet (FI) research [5][6]. New network architectural proposals depend on exhaustive tests before they are implemented in the real world. Thus, various experimental facilities, or testbeds, were developed [7][8]. Researchers, however, realized that interconnecting those testbeds is a requirement in order to carry out real experiments in geographically dispersed scenarios. This brings up many management challenges, because communication and access control agreements are necessary. The need to specify an IdM architecture in this context draws attention.

There are some proposals for federating FI testbeds. Among them, we highlight the Slice-Based Federation Architecture (SFA) [9], which is currently in use in testbeds such as OneLab, FIBRE Project (*Future Internet testbeds experimentation between BRazil and Europe*) [10], and PlanetLab. In SFA-based FI testbeds, users supply their credentials to get access authorization to a set of resources located in different institutions, such as a set of computers and a minimal specified bandwidth. Although SFA is the most important initiative to create a federation of FI testbeds, it presents open issues related to A&A. Briefly, this occurs because its proposal is focused on interconnecting resources through a *resource federation*. The A&A ends up in background, composed only of a simple authentication mechanism based on X.509 certificates and static profiles.

To illustrate our proposal a component architecture is shown in Figure 1. On top the identity federation is responsible for authenticating users. In the middle all components of a federated resource environment are depicted, with an attribute provider, an access control component, a credential translation

 ISBN: 978-1-61208-398-8

component (necessary to translate the credential from the identity federation to the testbed federation, which is presented in [11]), a service provider and the resource federation (i.e., FI testbeds or islands).

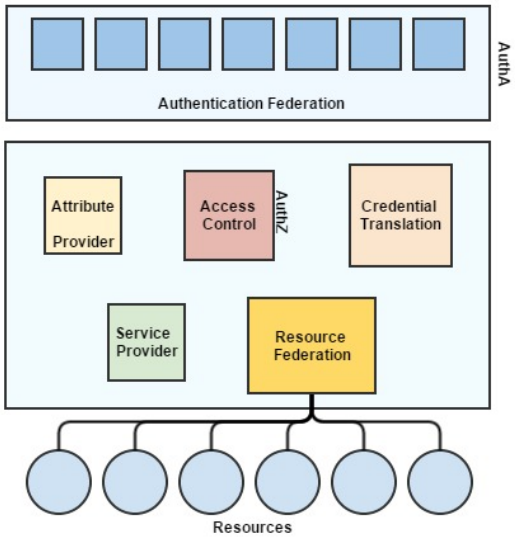


Figure 1. Proposed component architecture for authentication and authorization.

In this work, we propose a new authorization method for SFA-based testbeds. Our proposal integrates A&A federations based on Shibboleth [12] and a authorization framework based on Extensible Access Control Markup Language (XACML) [13]. Shibboleth implements the Security Assertion Markup Language (SAML) standard [14] and also supports Attribute-Based Access Control (ABAC), which has become a standard in 2014 [15]. Using ABAC, it is possible to implement more granular and dynamic access policies. Moreover, Shibboleth is used by the Brazilian academic federation, named CAFe, and also by eduGAIN. Our proposal allows the user to allocate resources in testbed federations based on attributes arising from an identity federation. We propose a generic framework for ABAC using an aggregate attribute mechanism that associates points for user attributes and resource attributes. Our goal is to use the access control proposal in the FIBRE testbed, which is an initiative of federated testbeds between Brazil and Europe, using CAFe and eduGAIN for authentication.

We implemented the proposed A&A architecture using a real experimentation laboratory called GIdLab. GIdLab provides a mirror of the CAFe federation, which serves as an experimental environment for new applications that use the federation. GIdLab also offers virtual machines, in which we configure some virtual testbeds and all ABAC infrastructure. This implementation allowed us to validate the proposal and to evaluate security features, comparing it to other proposals.

The rest of the article is organized as follows. Section II discusses related work. Section III shows an essential background of technologies and concepts necessary to understand the proposal. Section IV details our proposal and Section V presents current results. In Section VI, conclusions and future works are described.

II. Related Work

There are many initiatives to federate testbeds, such as GENI (Global Environment for Network Innovations), OneLab and FED4FIRE [5]–[8][16].

A recent initiative for creation of testbeds in Brazil and Europe is the FIBRE Project [10]. FIBRE proposes the construction of a network for large-scale experimentation, which includes wired and wireless environments, through the interconnection of small testbeds, called islands, in various parts of Brazil and Europe. Thus, FIBRE is strongly grounded in building a federated environment. Although these proposals are strongly based on the resource federation, using tools such as SFA, they present open issues related to identity management. These projects integrate testbeds using different control and management frameworks, each of them using a different user database and different authentication and access control methods. Tools such as SFA do not provide a proper A&A federation architecture to integrate such different environments.

Related work, in general, proposes the introduction of a standardized model (such as RBAC [17], ABAC [15]) for resource distributed environments such as grid computing, and more recently, cloud computing. The area where there is more related work is undoubtedly grid computating. We can cite works as [18][19], on which role-based access control models (RBAC) are applied. More recent works in cloud computing use ABAC for access control, such as [20]. However, we must emphasize the need to know how access control in these environments has been employed, but each distributed resource environment has its particularities.

In this paper, we present the first proposal for policy and Attribute-Based Access Control in FI testbeds, different of GENI ABAC proposal (the other – and only – similar proposal), when attributes are used to restrict and delegate access [21], introducing a new way to represent the resources, attributes and policies in this enviroment.

III. Background

This section presents an introduction of technologies and concepts needed to understand the solution proposed by this work.

A. SAML and Shibboleth

The Security Assertion Markup Language (SAML) standard [14] presents a set of specifications to define an infrastructure for dynamic exchange of security information between partners (e.g., institutions). SAML defines the roles of entities, “assertions” and transport protocols supported. Assertions use an XML format [22] for describing data, which represents the authorization of a user at a given time for instance. There are two main types of entities that compose an Authentication and Authorization (A&A) federation environment: Identity Provider (IdP is responsible for storing and providing information about users and their authentication and the Service Provider (SP) is responsible for offering one or more services (or resources). Shibboleth [12] implements SAML and allows

 ISBN: 978-1-61208-398-8

web applications to enjoy the facilities provided by the model of federated identity, such as the concept of Single Sign-On (SSO).

B. CAFe

Federated Academic Community (CAFe) is the Brazilian academic federation, encompassing education and research organizations. Through CAFe, a user keeps all his information at his home organization and can access services offered by institutions participating in the federation through SSO. CAFe uses standards and software solutions already available and adopted by other federations, such as Shibboleth. Besides maintaining the usual set of privacy policies, CAFe also comprises a set of tools for populating an Lightweight Directory Access Protocol (LDAP) repository with data from different corporate databases. Integrated to eduGAIN, the CAFe federation participates in the network of trust of GÉANT academic federation. In the FIBRE project context, CAFe is being proposed as the main means of authentication [11] for Brazilian users.

C. FIBRE project

The FIBRE project [10] is a partnership between Brazilian and European institutions in order to create a large-scale network virtualization testbed. Topologically, FIBRE can be seen as the union of a large European island and a large Brazilian island, which consists of several small islands, located in different universities and research centers.

In FIBRE, there are several control frameworks. To control OpenFlow equipment [23][6], the experimenter uses the OFELIA Control Framework (OCF) [24]. To control wireless equipment, FIBRE provides the Control and Management Framework (OMF) [25]. Moreover, FIBRE also has islands based on ProtoGENI, which is a control framework developed for the GENI project [26]. The idea is that FIBRE can provide different control interfaces and can aggregate an increasing number of islands.

D. XACML

XACML (eXtensible Access Control Markup Language) [13] is an XML-based standard language for declaring security policies by OASIS (*Organization for the Advancement of Structured Information Standards*), aiming at ensuring interoperability between authorization systems. Moreover, it is a language to declare access control policies, defining a format for request and response messages [13].

ABAC uses the XACML architecture, working with the same entities. For example, PEP (Policy Enforcement Point), PDP (Policy Decision Point) and PAP (Policy Administration Point), where PEP is responsible for translating requests and responses and PDP for deciding if any policy (defined in PAP) is applied.

E. Access Control Mechanisms

Access control (in this work, access control and authorization have the same meaning) is a fundamental mechanism for protecting a resource from unauthorized access or respecting security requirements. Specifically, an access control policy defines the conditions to which access to resources can be granted and to whom. With the increasing complexity of computing systems, access control methods have evolved from Mandatory Access Control (MAC) [27], Discretionary Access Control MAC (DAC) [28], Role-Based Access Control (RBAC) [17], to Attribute-Based Access Control (ABAC) [15]. In this work, ABAC is the focus of access control applied to resource federation for FI testbeds.

In [15], ABAC is an access control method where subject requests to perform operations on objects are granted or denied based on assigned attributes of the subject, assigned attributes of the object, environmental conditions, and a set of policies that are specified in terms of those attributes and conditions.

IV. ACROSS-FI Proposal

This section presents the ACROSS-FI proposal. At first, an attribute aggregator is proposed and validated, using an attribute provider to store application-specific user attributes. Then, a mechanism to generalize attribute values of both users and resources is introduced. Finally, the proposal of access control in the FI testbed environment is presented.

A. Attribute Aggregation

An attribute provider is necessary to store application-specific complementary attributes for a given user. Additional attributes are those employed only in a specific context, such as a trial project in networks. In the FI testbed scenario, many additional attributes can be necessary to access network resources, on the other hand, they are not necessary in other federation services. So, storing those additional attributes is a responsibility of the service provider (i.e., FI's testbeds), not the Identity Federation (e.g., CAFe).

Attribute aggregation models were studied [29][30], and in a nutshell, these papers introduce two models, when the user is responsible (behalf) to aggregate all distributed attributes (at different IdPs) or, alternatively, the SP is responsible to proceed with this aggregation. In this work we decided to develop a particular solution where the aggregation of attributes is implemented with the help of an attribute aggregator and one extra attribute provider, once our enviroment has particular characteristics as specifics attributes for specifics testbeds. We will see a similar approach on *attribute aggregation performed by linking service* presented by [29], when the SP is responsible to link all attribute's source. However, additional attributes stored in the attribute provider should not identify the user to which they are associated. Thus, a single and opaque ID was created in order to link the academic federation user with his extra attributes without identifying him, protecting him from malicious attacks. In this case, malicious attack can be a user modifying its attributes (or of other user) to obtain more privileges than he really should. With an opaque ID, this weakness is solved.

The attribute aggregation proposal is shown in Figure 2, where steps 1 and 2 are the process of user authentication at his home identity provider (IdP) in CAFe federation. Then, CAFe returns the user attributes in step 3 to the service provider (SP), which in turn forwards them to the attribute aggregator. An opaque attribute is sent to the attribute provider that returns only the user additional attributes without the knowledge of

 ISBN: 978-1-61208-398-8

which user that opaque attribute refers to, as shown in step 5. Additional attributes are kept in a local LDAP (IdP of Attribute Provider). In the end, the Attribute Aggregator gathers all attributes.

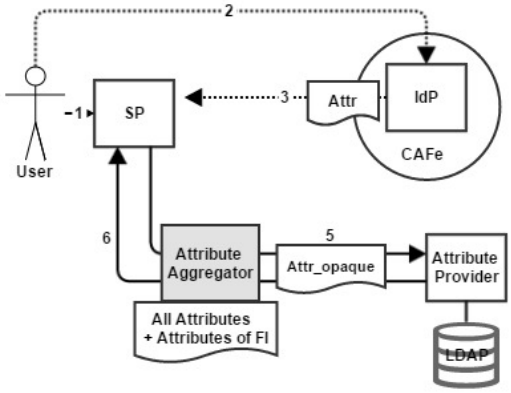


Figure 2. Attribute Aggregation Proposal.

The following equations show how the unique and opaque attribute to identify each user is generated:

$$\delta \leftarrow Attr_u(uid) \cup Attr_u(uidNumber) \quad (1)$$

$$Attr_U(opaque) \leftarrow hash(\delta) \quad (2)$$

Figure 3. Results of attribute aggregation process.

As an example for *uid* $\rightarrow$ *esilva@uff* with *uidNumber* $\rightarrow$ *1223*, concatenating these two attributes in an MD5 hash, we have the result $af2ec12ce73cc910358ddb400f4abb74$, which corresponds to the user ID at the Atribute Provider. It is noteworthy that only to validate the model a simple MD5 hash was used. Other modern cryptographic hashes, such as SHA-2, SHA-3, etc., should be used in a real environment.

Results of Attribute Aggregation. Briefly, the user proceeds with all steps involved in Shibboleth authentication. After this, Figure 3 shows all user attributes, the ones that came from his home IdP and additional attributes that came from the Attribute Provider (highlighted in red). They refer to specific attributes of the FI environment, i.e., *Shib-fibre-userEnable* (if user is active in FI testbeds), *Shib-fibre-omfAdmin* (if user is an administrator of OMF testbeds) and the opaque attribute, *Attr_opaque* (the user ID at the Attribute Provider).

B. A Generic Access Control Based on Attribute Scores

This work proposes a new method to generalize attribute values of both users and resources. This generic approach is applied on ABAC scenarios. At first, attributes are associated to points and those points are summed to determine a user level. Attribute points are predetermined by a global administrator (the global administrator of the FI environment). Algorithm 4 explains the procedure of computing a user attribute score. All possible attributes are contained in a list called $All\_Attributes$, where each attribute has a weight $Weight(Atribute)$. A simple normalization of attribute scores is applied (forcing the score to range from 0 to 1).

Data: User attributes.
Result: Score (Total of points).
1 **for** *Attribute in All_Attributes* **do**
2 **if** *Attribute.content* $\subset$ *User_List[Attribute]* **then**
3 $Total \leftarrow Total + Attribute.Point * Weight(Atribute)$;
4 **end**
5 **end**

Figure 4. Attribute Score.

When a user is associated to a level, a generalization can be used, because the policy access control does not need to know exactly which attributes a user has. User levels are used to define global and local access control policies. Global access control policies are defined by a FI testbed federation administrator. Local policies are defined by a FI testbed island administrator.

To illustrate the proposal of generalization based on attribute scores, one example is given in Table I. In this example, the maximum value is equal to 80 and the minimum is 0, when normalized ranges between [0-1], as follows:

$$\left(z_i^k\right)_N = \frac{z_i^k - z_{min}^k}{z_{max}^k - z_{min}^k}$$

where the result is a normalized number assuming the max and min attribute scores and the computed user score z_i^k.

TABLE I. AN EXAMPLE OF ATTRIBUTE SCORE.

An example of attribute score.					
Attribute	**value**	**points**	**weight**	**score**	**normalized**
brEduAffiliationType	student	10	3	30	
omfAdmin	TRUE	10	2	20	
institution	uff	8	1	8	
			Total	**58**	**0.725**

As shown in Table I, that user has score of 0.725 points. In the proposed model, it is assumed that the global administrator

 ISBN: 978-1-61208-398-8

will determine a number of levels L and score thresholds for each level. Thus, for a score $l_i < X \leq l_{i+1}$, the user will be on N_i level, where $1 \leq i \leq L$. So, in Table II the global administrator sets 3 levels, where the example user is associated to level 2.

TABLE II. AN EXAMPLE OF LEVEL DEFINITION BASED ON SCORES.

Level Definition	
Score	**Level**
$0 \leq X \leq 0.5$	01
$0.5 < X \leq 0.75$	02
$0.75 < X \leq 1$	03

TABLE III. ACCESS CONTROL POLICIES BASED ON SCORES FOR VIRTUAL MACHINES.

Access Control policies based on scores for virtual machines	
VMs	**Level**
$0 \leq X \leq 5$	1
$0 \leq X \leq 15$	2
$0 \leq X \leq 20$	3

The island (local testbed resources in one institution) administrator sets how many resources a user under a certain level can request. For example, Table III shows that the example user can request up to 15 virtual machines.

V. ACROSS-FI VALIDATION

A. Scenario

In this section, we explain how ACROSS-FI modules are interconnected. Figure 5 shows all components involved, from the authentication to authorization. It is possible to see a blue box on the top, where the authentication process occurs. Red boxes represent the authorization, with attribute aggregation and ABAC solution. It is also possible to see an XML document configured by a global administrator, concerning the creation of user levels used in authorization (Tables I and II).

Our case study is the FIBRE project. Based on Figure 5, the main steps, taken beginning from the federated authentication up to authorization to use resources protected by distributed access policies, can be seen, considering both local and global policies.

Thus, in **step 1**, the user accesses the service provider (SP) that will forward (**step 2**) to authentication, either through CAFe or FIBRE federated LDAP. The FIBRE federated LDAP is a tree that interconnects both Brazilian and European institutions that do not integrate CAFe, enabling a federated access to other users participating in the project. Such steps are traditionally used to create an SAML session [14] from the user to the SP access, redirecting through WAYF (Where Are You From) to its home IdP to proceed the authentication and exchange the user attributes.

In **step 3**, the authenticated user requests the list of resources available to the SFA federation. In this step, SFA is responsible to communicate with the SM (Slice Manager) having a global view of all AMs (Aggregate Managers) (**step 4**), which have direct contact with the island testbed resources. Thus, available resources are listed by a type of XML files, called RSpecs (Resource Specification) and returned to the user in **step 5**. Thereafter, the user may request the resources.

Step 6 is responsible to send through the SP the attributes to attribute aggregator, and the attribute aggregator is responsible to generate an opaque attribute, which identifies the user in the attribute provider, so that additional attributes of IF can be recovered without identifying the user directly to the IdP CAFe federation (**steps 7 and 8**).

Then, the SP receives all attributes from the attribute aggregator, in **step 9**, and forwards these attributes and the RSpec (identifying the requested testbed resources) in **step 10** to the PEP. The PEP then computes the user attribute score indicating a user level (see Section IV-B). In **step 11**, the PEP performs the conversion of RSpec files and score generated to an XACML request.

In **step 12**, the XACML request generated is sent to each FIBRE island, (the island's PIP, **step 13**, will check for additional attributes – optional). Then, in **step 14**, XACML policies that the island administrator previously registered through the PAP are returned to the PDP (**step 15**). At that time, a global policy is also checked through **step 16** and policy-combining algorithms presented by XACML are used, returning to the PDP (**step 17**) the decision. In **step 18**, the response is returned to the PEP that converts to RSpec and forwards it to the SP (**step 19**), stating if the user can or cannot allocate the requested resources.

B. Results

To validate the proposal, the GIdLab experimental laboratory was used, where a mirror of CAFe is available and all other proposed components. All modules, including the attribute aggregator and credential translation were developed. Other necessary features, such as how the SP communicates with SFA federation and the ABAC control access were also implemented.

As we saw in Section IV-A, the attribute aggregator was proposed and validated. Similarly, credential translation was also discussed and implemented in [11]. The proposal of attribute scores presented in Section IV-B was validated at GIdLab using the Sun's XACML implementation and three different virtual machines, one for PEP and PDP and two others to simulate the testbeds. Thus, after checking the user level and the RSpec, a number of resources (e.g., VMs) were requested, and the island's PAP checks if the user is allowed to allocate that number of requested resources. When the XACML response was received by the PDP, a global policy was verified too, and only after these steps the user received an RSpec with available resources.

VI. CONCLUSION AND FUTURE WORK

This work is motivated by the need of access control mechanisms in environments for evaluation of Future Internet proposals. In this context, federations for authentication and authorization can be used to facilitate the shared use of testbed resources for researchers belonging from different institutions.

Our main contributions are: 1) attribute aggregation model proposed and validated; 2) an ABAC model based on user scores and levels to associate dynamically users and resources

 ISBN: 978-1-61208-398-8

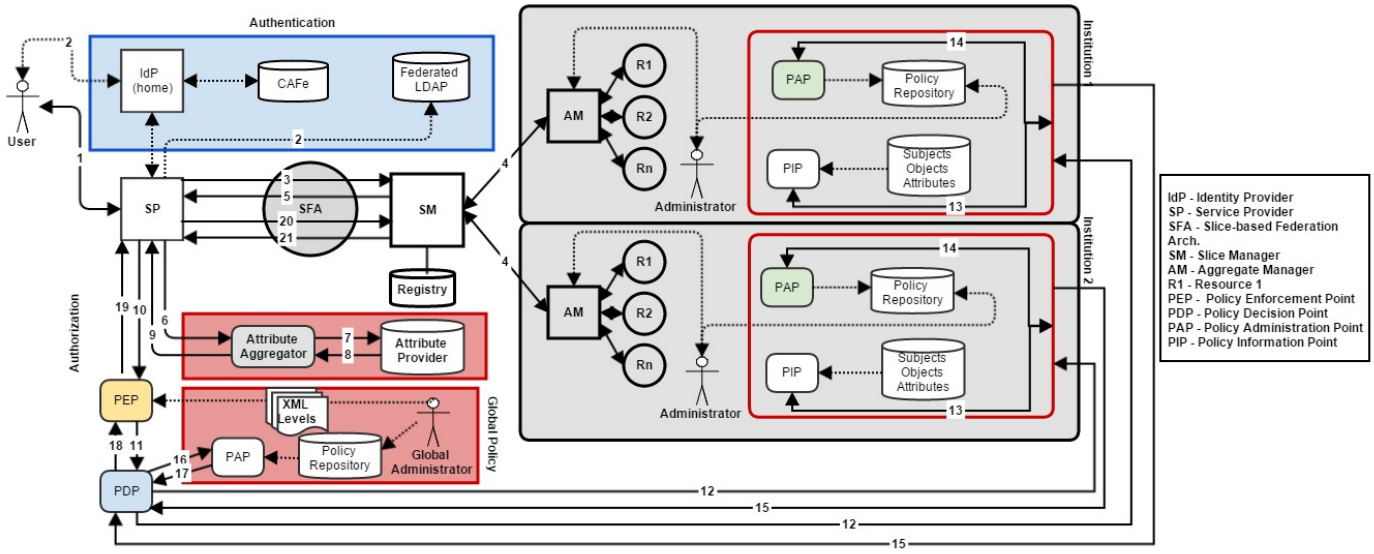


Figure 5. ACROSS-FI Scenario.

proposed and validated; 3) implementation of an integrated authentication and authorization solution for FI testbed environments validated in GIdLab.

As future work, we intend to generalize the proposed A&A solution to the concept of virtual organizations, where a subset of users from different home institutions may use services and share resources from other institutions. We also intend to develop configuration tools to facilitate the definition of attribute points and user levels and policies.

ACKNOWLEDGMENT

We thank RNP (GIdLab), CAPES, CNPq and FIBRE project for supporting this research.

REFERENCES

[1] J. Jensen, "Federated identity management challenges," in Availability, Reliability and Security (ARES), 2012 Seventh International Conference on, 2012, pp. 230–235.

[2] R. Dhungana, A. Mohammad, A. Sharma, and I. Schoen, "Identity management framework for cloud networking infrastructure," in Innovations in Information Technology (IIT), 2013 9th International Conference on, 2013, pp. 13–17.

[3] J. Leskinen, "Evaluation criteria for future identity management," in Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on, 2012, pp. 801–806.

[4] RNP, "CAFe - Federated Academic Community," http://portal.rnp.br/web/servicos/cafe-en, Feb. 2015.

[5] P. Stuckmann and R. Zimmermann, "European research on future internet design," Wireless Communications, IEEE, vol. 16, no. 5, 2009, pp. 14–22.

[6] R. Riggio, F. De Pellegrini, E. Salvadori, M. Gerola, and R. Corin, "Progressive virtual topology embedding in openflow networks," in Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on, 2013, pp. 1122–1128.

[7] S. Jeong and A. Bavier, "Geni federation scenarios and requirements," Tech. Rep., Jul. 2010.

[8] A. Falk, "Federation in geni - draft proposal - comments invited," in GENI Engineering Conferences - GEC11, Jul. 2011.

[9] L. Peterson, R. Ricci, A. Falk, and J. Chase, "Slice-based facility architecture," Tech. Rep., Jul. 2010.

[10] S. S. et. al, "FIBRE project: Brazil and Europe unite forces and testbeds for the Internet of the future," in Proceedings of TridentCom 2012, Jun. 2012.

[11] E. Silva, N. Fernandes, N. Rodriguez, and D. Muchaluat-Saade, "Credential translations in future internet testbeds federation," in 6th IEEE/IFIP Workshop on Management of the Future Internet (ManFI 2014)/NOMS, May 2014, pp. 1–6.

[12] T. Scavo and S. Cantor, "Shibboleth architecture," Tech. Rep., Jan. 2005. [Online]. Available: http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf

[13] T. Moses. eXtensible Access Control Markup Language TC v2.0 (XACML). OASIS. [Online]. Available: http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf [retrieved: Feb., 2005]

[14] OASIS, Security Assertion Markup Language (SAML) v2.0, Std., 2005.

[15] V. C. Hu, D. Ferraiolo, R. Kuhn, A. Schnitzer, K. Sandlin, R. Miller, and K. Scarfone, "Guide to attribute based access control (abac) definition and considerations," NIST Special Publication, vol. 800, 2014, p. 162.

[16] Future Internet Research & Experimentation, http://cordis.europa.eu/fp7/ict/fire/, March, 2015.

[17] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed nist standard for role-based access control," ACM Trans. Inf. Syst. Secur., vol. 4, no. 3, Aug. 2001, pp. 224–274. [Online]. Available: http://doi.acm.org.ez24.periodicos.capes.gov.br/10.1145/501978.501980

[18] B. et. al, "Identity Federation and Attribute-based Authorization through the Globus Toolkit, Shibboleth, Gridshib, and MyProxy," in 5th Annual PKI R&D Workshop, Apr. 2006.

[19] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations," Int. J. High Perform. Comput. Appl., vol. 15, no. 3, Aug. 2001, pp. 200–222.

[20] Y. Zhu, D. Huang, C. Hu, and X. Wang, "From rbac to abac: Constructing flexible data access control for cloud storage services," Services Computing, IEEE Transactions on, vol. PP, no. 99, 2014, pp. 1–1.

[21] M. Berman and M. Brinn, "Progress and challenges in worldwide federation of future internet and distributed cloud testbeds," in Science and Technology Conference (Modern Networking Technologies) (MoNeTeC), 2014 First International, Oct 2014, pp. 1–6.

[22] W3C, "Extensible Markup Language (XML) 1.0 (Fifth Edition)," Feb. 2015. [Online]. Available: http://www.w3.org/TR/REC-xml/

 ISBN: 978-1-61208-398-8

[23] M. et. al, “Openflow: enabling innovation in campus networks,” SIGCOMM Computer Communication Review, vol. 38, no. 2, Mar. 2008, pp. 69–74.

[24] A. Köpsel and H. Woesner, “Ofelia: pan-european test facility for openflow experimentation,” in Proceedings of the 4th European conference on Towards a service-based internet, ser. ServiceWave’11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 311–312.

[25] T. Rakotoarivelo, M. Ott, G. Jourjon, and I. Seskar, “OMF: a control and management framework for networking testbeds,” SIGOPS Oper. Syst. Rev., vol. 43, no. 4, Jan. 2010, pp. 54–59.

[26] ProtoGeni ClearingHouse, http://www.protogeni.net/wiki/ClearingHouse Desc, March, 2015.

[27] R. S. Sandhu, “Lattice-based access control models,” Computer, vol. 26, no. 11, Nov. 1993, pp. 9–19.

[28] R. Sandhu and P. Samarati, “Access control: principle and practice,” Communications Magazine, IEEE, vol. 32, no. 9, Sept 1994, pp. 40–48.

[29] D. Chadwick and G. Inman, “Attribute aggregation in federated identity management,” Computer, vol. 42, no. 5, May 2009, pp. 33–40.

[30] ——, “The trusted attribute aggregation service (TAAS) - providing an attribute aggregation layer for federated identity management,” in Availability, Reliability and Security (ARES), 2013 Eighth International Conference on, Sept 2013, pp. 285–290.

 ISBN: 978-1-61208-398-8

IVHM System Integration Network Performance Analysis using Different Middleware Technologies and Network Structure

Rajkumar Choudhary
Researcher, IVHM Centre
Cranfield University
Cranfield, Bedford, UK
e-mail: r.choudhary@cranfield.ac.uk

Suresh Perinpanayagam
Lecturer at Cranfield University
Cranfield University
Cranfield, Bedford, UK
e-mail:s.perinpanayagam@cranfield.ac.uk

Eugene Butans
Senior Research Fellow, SATM
Cranfield University
Cranfield, Bedford, UK
e-mail: eugene.butans@cranfield.ac.uk

***Abstract*—According to Aircraft Crashes Record Office (ACRO), total number of accidents occurred from 1999-2013 were 2556 in worldwide, were primarily due to loss-of-control in flight, controlled flight into terrain, and system/component failure. These accidents caused big capital loss for aircraft industry and 18987 deaths. Aircraft manufacturers are investing a huge amount of money to minimize these accidents by implementing new technologies, e.g., IVHM (Integrated Vehicle Health Management) in legacy and new generation aircraft. In aircraft industry, maintenance costs represent the third largest expense item after labor and fuel costs for both regional and national carriers with maintenance costs commonly comprising 15-18% of the operational expenses. By implementing IVHM technologies not only the maintenance costs can be reduced, also the fatality rate can be minimized. IVHM can provide more specific scheduled maintenance, onboard diagnostics and prognostics services. The aim of this paper is to investigate, about finding network architecture suitable for IVHM integration in vehicles (e.g., aircraft) that should be able to support interoperability between multiple vendors' IVHM components and insertion of new IVHM capabilities using simulation and optimization technique. To develop IVHM network architecture, essential data such as bandwidth, data rate, throughput, latency and performance in communication network will be collected using various enabling technologies (i.e., middleware) and OSA-CBM (Open System Architecture for Condition Based Monitoring) data model. Using simulation tools, e.g., OPNET (Optimized Network Engineering Tools), these sample data will be tested at large scale environment (e.g., aircraft or train). After simulation, multi-objective optimization will be used in trade-off analysis that aims to find cost effective and fully functional IVHM network architecture.**

Keywords-Distributed systems; IVHM; systems integration; architecture; OSA-CBM; middleware

I. INTRODUCTION

According to Boeing, an airplane has an average economic life of about 27.2 years. [1] This life span can be further divided into three categories: airplane useful life, airplane economic life, and airplane service life. In maintenance and service, the aviation industry invest enormous amount of money, to keep the aircraft operationally available throughout in-service period of its life cycle. The IVHM emerges as an advanced diagnostics and prognostics techniques provider. IVHM technologies can impact in both acquisition cost and maintenance cost throughout the life cycle of aircraft by efficiently identifying and predicting failures to enable the effective planning of maintenance tasks [2].

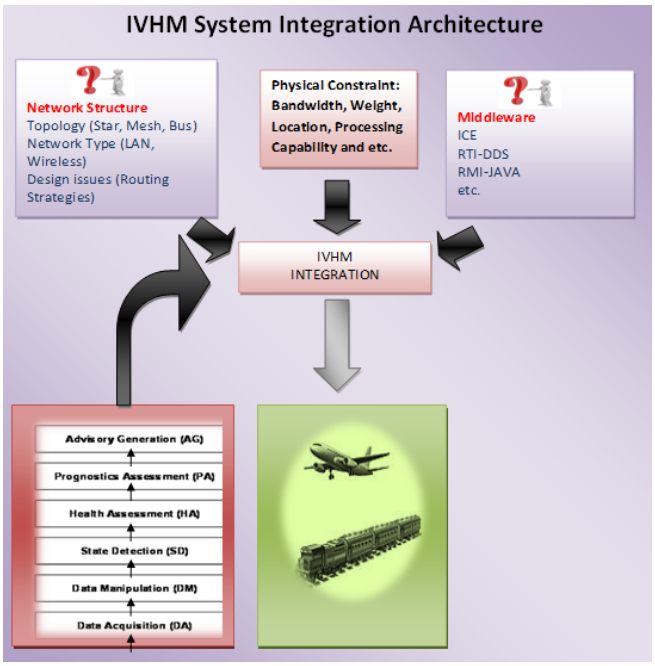


Figure 1. IVHM System Integration Architecture

This research is focused on Designing and Development of a Large-Scale IVHM Network Architecture using Multi-Objective Simulation Optimization. The architecture should be flexible and extensible that can directly support the implementation of upcoming

 ISBN: 978-1-61208-398-8

technologies of IVHM and various IVHM cooperate partners. Extensive research has been conducted by Boeing, Thales and other system integrators for system integration of IVHM. Most of the work has been done for development of individual subsystem and integration of IVHM on the aircraft. An adaptable architecture is required which can support interoperability between multiple vendors' IVHM components and insertion of new IVHM capabilities [2]. In the past, there have been very limited studies done on different approaches in integration of IVHM components. Aircraft industry is looking for scalable, flexible, economical and reliable network architecture for IVHM technologies integration in aircraft. Planning the reliability issues is necessary such as reliability of communication networks and reliability of Network costs. The penalties issue needs to be considered which may exceed the profits from the providing IVHM integration services. How to integrate this IVHM architecture in an efficient and cost-effective way to maximize the overall product rate is a challenge to the IVHM industry [3]. IVHM systems include sensors, processing units and software, which comes from multiple suppliers having different configurations. Several IVHM based projects are in progress for different technologies using various programming languages and contrary platforms by different organizations, for example: Integrated Intelligent Flight Deck (IIFD) Project and Aircraft Aging and Durability (AAD) Project. IVHM technologies can be developed following the OSA-CBM data model as a common standard. OSA-CBM (Open Source Architecture for Condition Based Monitoring) is developed by MIMOSA, which is based on ISO standards (i.e., ISO13374-1, ISO13374-2 and ISO13374-3).

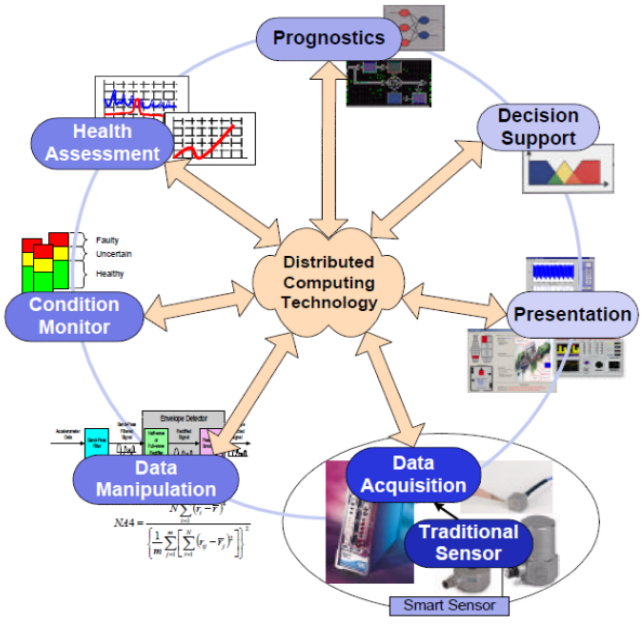


Figure 2. OSA-CBM Model developed by MIMOSA (Machinery Information Management Open Systems Alliance) [4]

As shown in Figure 2, OSA-CBM model consists of seven layers, which are Data Acquisition (DA), Data Manipulation (DM), State Detection (SD), Health Assessment (HA), Prognostics Assessment (PA), Advisory Generation (AG) and human interface layer.

All OSA-CBM based IVHM systems follow the same pattern. Various sensors are used to collect data in DA layer, which is later transformed into a suitable format in DM layer, and this data is then analyzed using knowledge discovery algorithms. The data from lower layers is used to know the current state of component in SD layer. In HA layer, the current health of the component is analyzed based on data collected from previous layers. The RUL (Remaining Useful Life) of subsystem and prognostics details are saved and can be assessed in PA layer. Later this data is required to be shared across maintenance departments to take appropriate actions such as arrangement of spare parts if problem in any component has been detected. The AG layer sends information to relevant department to take appropriate actions if any fault detected. Lastly, human interface layer provides information to access data in OSA-CBM model. The IVHM practitioners are facing difficulties with determining the best method for interconnecting the system's components via communication networks. There is a need to tackle many issues for building large network architecture such as bandwidth saturation events (Point in which all available bandwidth is used up), broadcast impact on CPU processing, unpredictable reachability, address collision, unused duplicate circuits, non-optimal routing, limited network management, isolated configuration control and support a multi-protocol environment [5].

The addition of IVHM capabilities shall be further explained in Section II. The network architecture requirements related to IVHM technology insertion and an open systems approach to systems integration are covered in Section III. Section IV covers the modeling to evaluate the network performance of different IVHM implementations. The result of experiments is discussed in Section V. Lastly, the conclusion is covered in Section VI.

II. INSERSION OF IVHM CAPABILITIES

A typical IVHM system involves many different components that vary in bandwidth demand. IVHM system needs to be able to support multiple types of interconnection networks for hardware and software components that are dramatically different in their routing capabilities. Study of information exchange between the different subsystems and the system level is highly recommended for insertion of IVHM capabilities, which is essential for communication issues, synchronization, and input/output functionality [6]. The IVHM network architecture must provide a methodology for adding future network technologies without affecting existing IVHM components and interoperability across existing interconnection networks. The IVHM is facing difficulty with determining the best

 ISBN: 978-1-61208-398-8

way for interconnecting its components via communication networks. There is a need to tackle many issues for building large network architecture such as bandwidth saturation events, broadcast impact on CPU processing, unpredictable reachability, address collision, unused duplicate circuits, non-optimal routing, limited network management, isolated configuration control and support a multi-protocol environment [5]. The physical constraints of IVHM architecture are Acquisition of high fidelity data, cost of certification, limited communication Bandwidth, limited local processing, post flight test maturation, size, weight and power [7]. In order to overcome these constraints the IVHM architecture should incorporate various features such as Health ready subsystems (e.g., generator (IDG (Integrated Drive Generators)), open system (e.g., OSACBM), hierarchical-and-Distributed, partitioning of flight and Enabling technologies (e.g., Chafing protection system) [8]. An IVHM system is more than just a set of IVHM technologies. The technologies must work together in a realistic environment and must provide significant safety improvements to justify the development, integration, and costs associated with these technologies [6].

III. ARCHITECTURE REQUIREMENTS

A. *Available middleware technogies for implementation of IVHM Systems*

Middleware-enabling technologies are now used for distributed real-time and embedded (DRE) systems that control communication among devices in physical, chemical, biological, or defense industries [8]. Middleware that can satisfy stringent quality of service (QoS) requirements, such as predictability, latency, efficiency, scalability, dependability and security, can be used for development of IVHM network architecture.

TABLE I ENABLING TECHNOLOGIES –MIDDLEWARE [9]

Distributed Data Model	Middleware	OSA-CBM		Mapping Language	Protocol	Data Bus	Standard
		Interface	Data Model				
Client/Server	CORBA (TAO/JacORB)	Yes	Valuetype (from CORBA ver.2.3)	IDL	UDP (TAO), TCP	Ethernet, AFDX	OMG's CORBA
	ICE	Yes	class	Slice	UDP, TCP	Ethernet, AFDX	Proprietary
Publish/Subscribe	IceStrom	Yes	class	Slice	UDP, TCP	Ethernet, AFDX	Proprietary
Data Distribution service	RTI DDS	No	valuetype (Proprietary Extension)	IDL, XML	UDP, TCP	Ethernet	OMG's DDS-DCPS
	OpenSplice	RMI Extension	valuetype (DLRL)	IDL	UDP, TCP	Ethernet	OMG's DDS-DCPS /DLRL

Acronyms: AFDX –Avionic Full-Duplex Switched Ethernet, DCPS –Data Centric Publish/Subscribe, DDS –Data Distribution Service, RMI –Remote Method Invocation, TAO –C++ Implementation of CORBA, TCP –Transmission Control Protocol, DLRL –Data Local Reconstruction Layer, IDL –Interface Definition Language, JacORB–Java Implementation of CORBA, OMG –Object Management Group, UDP –User Datagram Protocol, XML –extensible Markup Language

In Table 1, a detailed comparison of important distributed computing object middleware technologies, which include ICE, IceStrom, RTI DDS, OpenSplice and CORBA on the basis of mapping language, protocol, data bus, standard, and OSA-CBM framework support, has been done [9].

IVHM subsystems can be implemented using any of these distributed data models: Client/Server, Publish/Subscribe and Data Distribution service. Different middleware supports different Distributed Data Model, e.g., ICE middleware supports Client/Server. All middleware use some kind of mapping language which provides interoperability. TCP and UDP are the most common protocols. CORBA, ICE and IceStorm can work on both Ethernet and AFDX data bus whereas RTI DDS and OpenSplice support only Ethernet.

Client/Server model utilize OSA-CBM interfaces and supports the required data flow characteristics (i.e., trigger by state changes or push all). It has an advantage of simplicity but the disadvantage is its direct connection between IVHM subsystems (i.e., tightly coupled).

Publish/Subscribe model uses Publish/Subscribe server as a global data space which manages data subscriptions and responsible for data delivery. It is loosely coupled as there is no direct connection between IVHM components. It has some disadvantages such as single point of failure and data latency (i.e., Data takes longer to deliver).

Data Distribution service combines various Publish/Subscribe server. Localized Publish/Subscribe server carries out data delivery. The service named dynamic discovery is used to find out IVHM components and what data are available for them. Removing or upgrading a subsystem will not affect other subsystems.

B. *OSA-CBM and its interfaces to implement IVHM systems*

The OSA-CBM specification is a standard architecture for moving information in a condition-based maintenance system. A more in depth look reveals a way to reduce costs, improve interoperability, increase competition, incorporate design changes, and further cooperation in the realm of condition-based maintenance [10]. Open system IVHM architecture can be built using OSA-CBM standard. OSA-CBM focuses on IVHM functionality and associated data models. Most OSA-CBM works are from Boeing, GE Aviation and ARL. Research need to be done on how OSA-CBM can be implemented or how IVHM architecture can be designed to compliant with OSA-CBM. An example of communication between two IVHM components using RMI-Java middleware and OSA-CBM data model is shown in Figure 3.

 ISBN: 978-1-61208-398-8

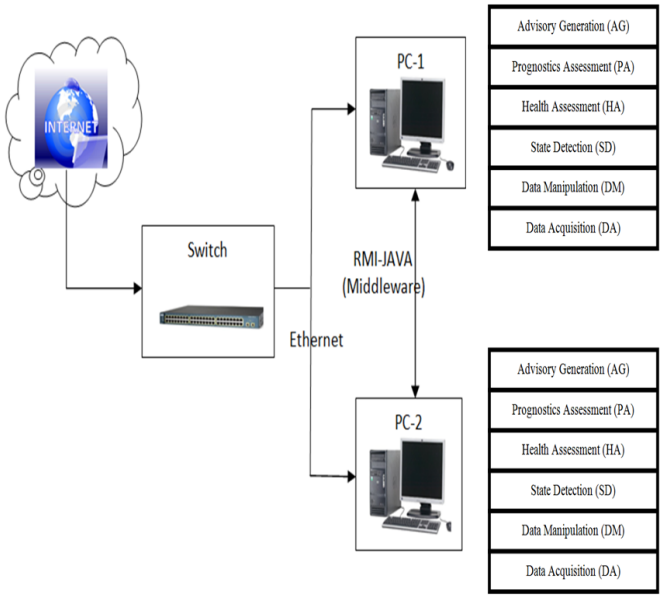


Figure 3. Communication between two IVHM Components using RMI-Java middleware and OSA-CBM data model

IV. MODELLING AND EXPERIMENTS

A. *RMI-Java Latency performance test:*

Hardware and software used for experiment:

PC1 CONFIGURATION : Intel Pentium CPU 2127U @ 1.90GHz, RAM – 4 GB, Windows 8.1 – 64bit Operating system
PC2 CONFIGURATION : Intel Pentium CPU 2127U @ 1.90GHz, RAM – 4 GB, Windows 8.1 – 64bit Operating system
Software used : Netbeans IDE 8.0.1
Java version : JDK 1.7
Middleware : RMI-Java
Data Model used : OSA-CBM
Overhead included from : OSA-CBM, RMI-Java Middleware and etc.

For experimental setup, two PCs (Personal Computers) are used to create a Client/Server model. Server acts as a service provider, and a client that acts as a service receiver. OSA-CBM is implemented in both as a common data model.

TABLE II RMI-JAVA LATENCY TEST

Data Transfer Size (bytes)	Latency per Data Transfer (milliseconds)
8	0.01969
16	0.03328
32	0.06769
64	0.12925
128	0.25379
256	0.51476
512	1.06714
1024	2.0008
2048	4.1136
4096	8.8302
8192	17.458

In Table II, the intention is to find the performance of RMI-Java using OSA-CBM data model by increasing the size of data.

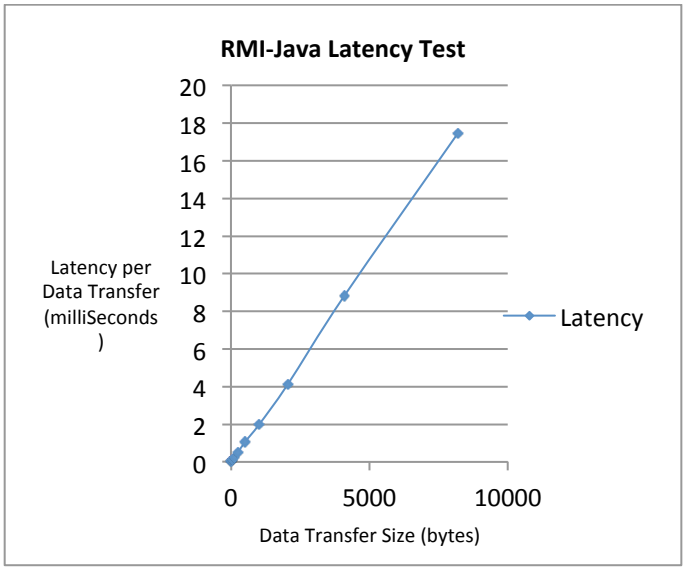


Figure 4. RMI-JAVA Latency test

Figure 4 illustrates the performance test of RMI-Java, based on latency and data size.

B. *RMI-Java Throughput performance test:*

Throughput (Kbit/second) = (Latency per data transfer (ms)/1000) *(data transfer size (bytes)/1000) * 8

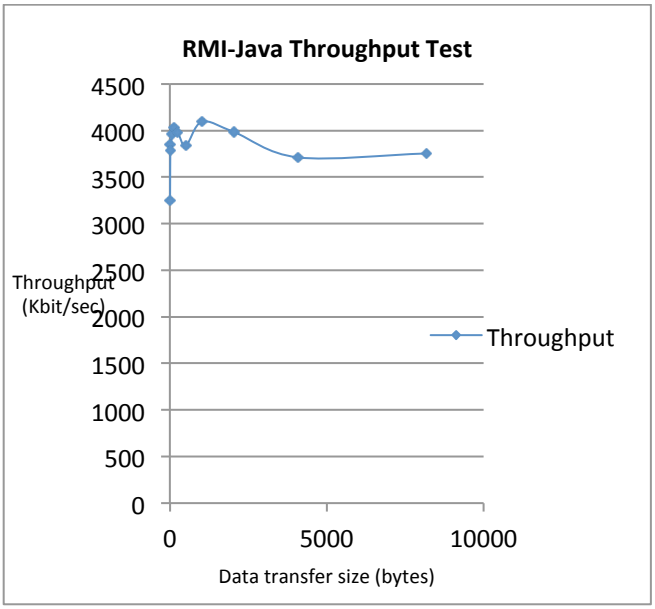


Figure 5. RMI-JAVA Throughput test

In this section, the throughput is calculated using values from Table II and the result is shown in Figure 5.

V. RESULTS AND DISCUSSION

As the work in progress, the result shows the performance of one middleware which is RMI-JAVA. RMI-JAVA shows low and predictable latency that scales linearly with message size. The performance test on ICE, RTI-DDS will also need to be done. The result aims to find the best system integration network architecture for IVHM. IVHM brings more sophisticated condition based maintenance system and problems at the same time, to synchronize with existing components and adapt with upcoming upgrades of the components. And, the components belong from different vendors with different configurations. After comparing several different enabling technologies (e.g., RTI-DDS middleware, ICE, CORBA and IceStrom), different mode of networks (e.g., LAN, WAN and wireless), different distributed data mechanism (e.g., Client/Server, Publish/Subscribe and Data Distribution service), different protocols (e.g., UDP and TCP/IP) and different network topologies (e.g., Star. Mesh and Ad-Hoc), this project will focus on implementing IVHM technologies on aircraft (vehicle) using best open system distributed communication system.

VI. SUMMARY AND FUTURE WORK

In summary, this project is about simulating IVHM Architecture and optimizing related parameters to find out a suitable way of system integration. The literature review can be generalized in six respects which are Aerospace operation and maintenance, IVHM, System integration, Network embedded, modeling and simulation at large scale platform (vehicle) network and simulation optimization. The research problem is to find out a suitable IVHM System integration strategy and how the strategy influence the QOS and cost to the IVHM systems which will be used in the IVHM implementation in vehicles(e.g., aircraft). Main data of the model, such as type of Data and network requirements of the related IVHM subsystems will be collected in the later months. Data characteristics (e.g., Bandwidth requirement, data speed and data security) will be collected using OSA-CBM model with different middleware technologies (e.g., ICE and RTI-DDS). An automated design environment will be created which can be used to find best possible network for IVHM integration.

REFERENCES

[1] Helen Jiang. (2013, Mar.) aircraft_economic_life_whitepaper.pdf. [Online], Accessed: [2014,Dec] http://www.boeing.com/assets/pdf/commercial/aircraft_economic_life_whitepaper.pdf

[2] T. Sreenuch, "PnP IVHM Architecture IVHM Seminar," , Cranfield University, Cranfield, 2010.

[3] Kirby Keller, T Dave Wiegand, Kevin Swearingen, Chris Reisig, Alan Gillis, MikeVandernoot, , "An architecture to implement Integrated Vehicle Health Management systems," *IEEE*, vol. 1, pp. 2-6, 2001.

[4] MIMOSA. MIMOSA. [Online], Accessed: [2015,Jan] http://www.mimosa.org/?q=node/350

[5] JeffryS.Yaplee and Robert 0.Denney "Building a Large Network at the Boeing Company," *IEEE*, vol. 1, pp. 20-22, 1988.

[6] Principal Investigator: Ashok N. Srivastava, Ph.D. Project Scientist: Robert W. Mah, Ph.D. Project Manager: Claudia Meyer, (2009) National Aeronautics and Space Administration. [Online], Accessed: [2015,Jan] http://www.aeronautics.nasa.gov/nra_pdf/ivhm_tech_plan_c1.pdf;

[7] Kirby Keller, Jim Peck, Kevin Swearingen, and Dan Gilbertson, "Architecture for Affordable Health Management," *AIAA*, vol. 1, pp. 1-3, 2010.

[8] Nanbor Wang. Douglas C. Schmidt, and. Aniruddha Gokhale (2014, Aug.) QoS4DRE.pdf. [Online], Accessed: [2014,Dec] http://www.cs.wustl.edu/~schmidt/PDF/QoS4DRE.pdf

[9] T. Sreenuch, "PnP IVHM Architecture," in *IVHM Seminar*, Cranfield University Cranfield, 2012, p. 10.

[10] MIMOSA. [Online], Accessed: [2015.Jan] http://www.mimosa.org/?q=resources/specs/osa-cbm-330

[11] OptTek Systems Inc. [Online], Accessed: [2015,Jan] http://www.opttek.com/OptQuest

[12] Jacob H. Christensen ,David B. Anderson, Bryan D. Hansen "Scalable Network Approach for the Space Plug-and-play Architecture," *IEEE*, vol. 6, 2012.

[13] Mostafa Fazeli, "Assessment of throughput under opnet modeler simulation tools in mobile ad hoc networks(MANETs)," *IEEE*, 2011.

[14] T. Sreenuch, A. Tsourdos I.K Jennions, "Distributed embedded condition monitoring systems based on OSA-CBM standard," *CSI*, pp. 238-246, 2012.

[15] Sumair Khan, "Performance Comparison of ICE, HORB, CORBA and Dot," *International Journal of Computer Applications*, vol. 3, 2010.

[16] Kalyan Perumalla, Alfred Park, Hao Wu, Mostafa H. Ammar Richard M. Fujimoto. Large-Scale Network Simulation: How Big? How Fast? [Online], Accessed: [2014,Dec]

 ISBN: 978-1-61208-398-8

http://www.cs.mcgill.ca/~carl/largescalenetsim.pdf

[17] Shahrul Kamaruddin Rosmaini Ahmad, "An overview of time-based and condition-based maintenance in industrial application, Pages 135–149," in *Computers & Industrial Engineering*.: Computers & Industrial Engineering, August 2012, vol. 63. [Online], Accessed: [2015,Jan]. http://www.sciencedirect.com/science/article/pii/S0360835212000484

[18] Richard E. Schantz and Douglas C. Schmidt, "Middleware for Distributed Systems," vol. 1, no. 1. [Online], Accessed: [2014,Dec] https://www.dre.vanderbilt.edu/~schmidt/PDF/middleware-encyclopedia.pdf

[19] T Sreenuch, "IVHM CONOPS," in *IVHM Seminar*, Cranfield University Cranfield, 2012.

[20] Grant A. Gordon, Honeywell Laboratories Dimitry Gorinevsky, "Design of Integrated SHM System for," vol. 1, no. 1, 2005.

 ISBN: 978-1-61208-398-8

A Lightweight Approach to Manifesting Responsible Parties for TCP Packet Loss

Guang Cheng
School of Computer Science, Southeast University, P.R.China
Key Laboratory of Computer Network &
Information Integration, Ministry of Education, P.R.China
email: gcheng@njnet.edu

Yongning Tang
School of Information Technology
Illinois State University, USA
email: ytang, tbgyires@ilstu.edu

Tibor Gyires

***Abstract*—Troubleshooting TCP packet loss is a crucial problem for many network applications. TCP packets could be lost in different network segments for various reasons. Understanding the responsible parties for TCP loss is an important step for network operators to diagnose related problem. However, TCP is designed for end-to-end control. It is difficult for any third party to detect whether and where (even coarsely) TCP packet loss has occurred. We design *TCPBisector*, a lightweight efficient diagnosis tool to manifest responsible parties for TCP packet loss. *TCPBisector* divides the responsibility between "My" and "Other" parties or networks (denoted as Net_m and Net_o) conceptually delimited by a passive network measurement point (denoted as MP), and quantifies the responsibility by using TCP packet loss ratios on the corresponding networks. The evaluation shows that the *TCPBisector* can accurately estimate TCP packet loss ratios with estimation error rate 3.5-6.9%.**

Keywords- responsibility; performance diagnosis.

I. INTRODUCTION

TCP packets could be lost in different network segments for various reasons, including network congestion, packet corruption, faulty network components, and network misconfiguration. Understanding the responsible parties for TCP packet loss is an important step for network operators to troubleshoot the problem. However, TCP [17] is designed as an end-to-end control protocol. It is difficult for any third party to detect whether and where (even coarsely) TCP packet loss has occurred.

TCP performance monitoring and diagnosis have been extensively studied. Several sophisticated network monitoring frameworks [7][18] and intrusive active probing techniques [1][8][9] were proposed to pinpoint the root cause of TCP packet loss. Many previous work [2]-[6] also focused on comprehensively understanding TCP behaviors (e.g., including TCP window sizes, TCP retransmission and reordering) and its correlation with the actual network performance (e.g., network throughput and congestion). Maintaining accurate and complete TCP flow information is critical for this type of study, which typically requires large memory space and high computing power. Recent study [16] focused on providing real time TCP monitoring and performance diagnosis based on various flow sampling techniques, which may skip important flow information.

In this paper, we propose *TCPBisector*, a lightweight tool to help network operators to answer one critical question: "How much should I (or other parties) be responsible for TCP packet loss?"

As shown in Figure 1, *TCPBisector* divides TCP packet loss responsibility between "My" and "Other" parties or networks (denoted as Net_m and Net_o) conceptually delimited by a passive network measurement instrument (denoted as Measurement Point or MP), and quantifies their responsibilities by using TCP packet loss ratios on "My" and "Other" networks (denoted as LR_m and LR_o), respectively.

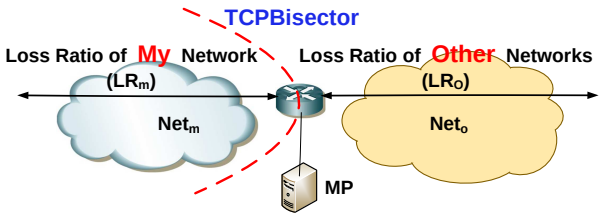


FIGURE 1: *TCPBisector*: a tool to bisecting responsible parties for TCP Packet Loss between Net_m and Net_o

Our work makes two contributions. Our first contribution is TCP behavior modeling. In the paper, we show that TCP presents different behaviors at the MP when TCP packets are lost by different parties between Net_m and Net_o. Accordingly, we model TCP behaviors by using several easy-to-track network events that allow the MP to ascribe the TCP packet loss responsibility to different parties.

Our second contribution is an efficient TCP packet loss inference algorithm. Instead of studying the causality of TCP packet loss, the *TCPBisector* only requires a small set of essential TCP related events commonly observable in various TCP packet loss scenarios. Thus, the *TCPBisector* can be used effectively and efficiently to infer occurred TCP packet loss and further identify their relative occurring locations, without suffering from the overhead of distinguishing TCP loss scenarios as shown in many related work [5][15][16]. Our inference algorithm presents the computation complexity of $O(n)$ and only requires a bounded memory space.

TCPBisector requires only one passive network measurement instrument (i.e., MP) as shown in Figure 1. The MP can be deployed arbitrarily on network depending on how

 ISBN: 978-1-61208-398-8

a responsibility scope defined under different monitoring strategies. Essentially, Net_m represents the scope of "my" responsibility, and Net_o shows the boundary of others. Depending on the different deployment strategy, the Net_m can be an enterprise network using cloud services, or a data center providing cloud services.

The rest of the paper is organized as the following. Section II describes the related work, and Section III introduces the TCP behavior modeling. Section IV presents the modularized *TCPBisector* as a system and discusses the algorithm for inferring LR_m and LR_o. We show the validation of our system via both emulations and experiments on a Tier-1 network in Section V. Section VI concludes our work.

II. Related Work

Numerous measurement studies have investigated the characteristics of TCP connections in the Internet, either via actively measured end-to-end properties (e.g., loss, delay, throughput) of TCP connections, or passively characterized a connection's aggregate properties (e.g., size, duration, throughput). Various TCP measurement methodologies and metrics have also been proposed [10]-[13] to monitor TCP performance via a set of important TCP parameters (e.g., RTT[14]).

Among various TCP related parameters, TCP packet loss is one of the most important metrics. Many scholars have proposed a variety of methods for TCP packet loss ratio estimation. Sommers et al., [1] proposed to send probe packets by the sender, and view the number of probe packets at the receiver that arrives to estimate end-to-end packet loss ratio. Benko and Veres [2] proposed to use the observed TCP sequence numbers to estimate TCP packet loss. Ohta and Miyazaki [3] explored a passive monitoring technique for packet loss estimation relying on hash-based packet identification. Friedl et al. [4] compared flows with sender and receiver for computing the packet loss,

Jaiswal et al. [5][15] presented a passive measurement methodology that observes the TCP packets in a TCP connection and infers/tracks the time evolution of two critical sender variables: the sender's congestion window (cwnd) and the connection round trip time (RTT). Allman et al. [6] estimated the packet loss ratio by observing the sender's retransmit packets. Nguyen et al. [7] built a model called the HSMM to analyze the packet loss. Zhang et al. [8] analyzed packet loss, delay and bandwidth from the random packet of the entrance and packet loss. *STA* [18] developed an efficient packet classification technique which is used to infer the loss and reorder rates of individual TCP flows.

Recent research has studied how to diagnose TCP performance issues in clouds. Ghasemi et al. [16] proposed a heuristic inference algorithm to infer several important TCP parameters (e.g., congestion-window size and the TCP state) from sampled TCP related statistics (e.g., RTT).

TCPBisector proposed in this paper is to coarsely bisect TCP packet loss responsibility between interior and exterior networks. *TCPBisector* is designed based on the fact that observable TCP behaviors could be different on different network segments along the same end-to-end path under the same network condition. *TCPBisector* aims at providing a practical, lightweight, and real-time tool for both cloud users and service providers understand network conditions between Net_m and Net_o.

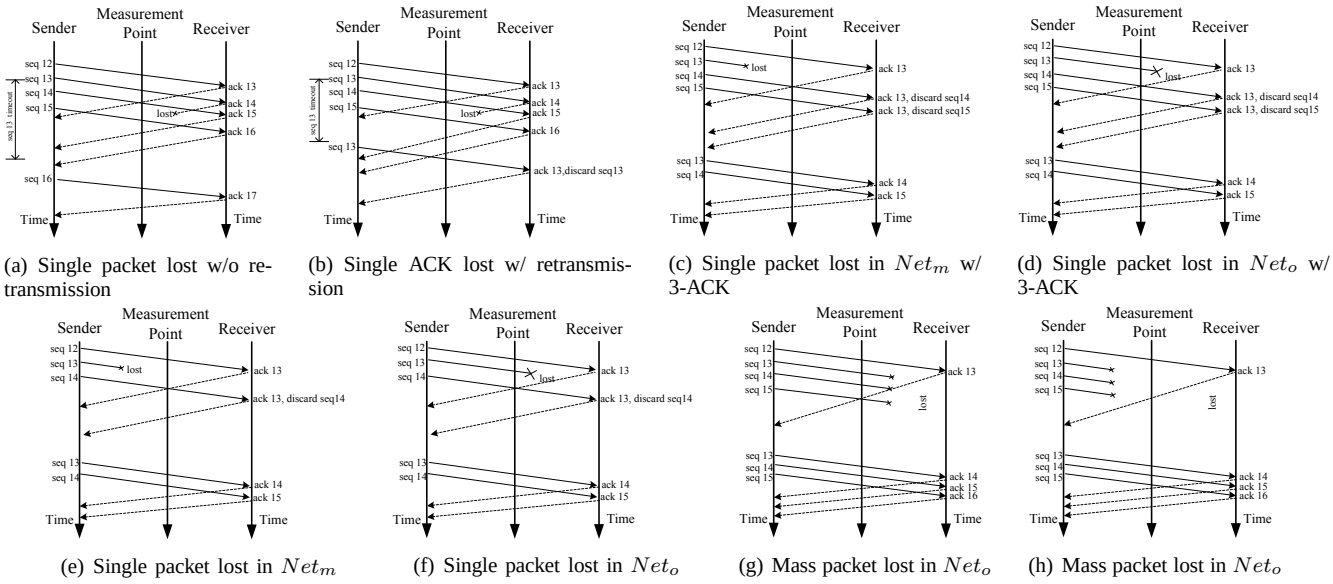
III. TCP Behavior Modeling

Although TCP is designed as an end-to-end control protocol, we show that the MP in the middle still can discern differences on the corresponding TCP behaviors when packet loss occurred on the different responsible parties (i.e., Net_m and Net_o). In the following, we will first illustrate several representative TCP packet loss scenarios. Then we will define two TCP behaviors distinguishable by the MP so as to ascribe the TCP packet loss responsibility to Net_m or Net_o.

A. TCP Loss Scenarios

TCP behaves differently in response to varying network condition. More importantly, TCP presents different observable behaviors at the MP when TCP packets loss occurred in Net_m or in Net_o. In the following, we illustrate our ideas via several representative TCP packet loss scenarios as shown in Figure 2. In all the scenarios, we assume the sender is from Net_m and the receiver is located within Net_o.

- ACK loss: Figure 2(a) & 2(b) show that a data packet from the sender has successfully delivered to the receiver. However, one acknowledgement packet (i.e., ACK_{14}) from the receiver was lost. In the scenario shown in Figure 2(a), since the following ACKs (i.e., ACK_{14} and ACK_{15}) arrived before a retransmission timeout event triggered at the sender, no retransmission occurred. Otherwise, the sender retransmitted the unacknowledged packet (i.e., Seq_{13}) as shown in Figure 2(b). Apparently, it appeared to the MP as if no packet loss in the first loss scenario (shown in Figure 2(a)). In this scenario, only one ACK lost in Net_o before passing the MP. However, the following ACKs successfully arrived at the sender, which took over the responsibility of the lost ACK. Thus, considering this scenario the same as no TCP loss makes sense practically. In the scenario shown in Figure 2(b), the MP could observe the occurrence of data retransmission.
- Single packet loss with 3-ACK: Figure 2(c) and Figure 2(d) show a type of common TCP loss scenarios, in which one data packet (i.e., Seq_{13}) was lost. Consequently, the sender received three duplicate ACKs (denoted as 3-ACK). Depending on where the data packet lost, the MP may only observe three consecutive ACKs as shown in Figure 2(c) if the loss occurred in Net_m; or observed duplicated data



(a) Single packet lost w/o retransmission

(b) Single ACK lost w/ retransmission

(c) Single packet lost in Net_m w/ 3-ACK

(d) Single packet lost in Net_o w/ 3-ACK

(e) Single packet lost in Net_m

(f) Single packet lost in Net_o

(g) Mass packet lost in Net_o

(h) Mass packet lost in Net_o

FIGURE 2: The Illustration of Representative TCP Packet Loss Scenarios

packets in addition to three consecutive ACKs as shown in Figure 2(d) if the loss occurred in Net_o.

- Single packet loss with timeout: Figure 2(e) and Figure 2(f) show another type of common TCP loss scenarios, in which the sender didn't receive three duplicate ACKs. Instead, a timeout event was triggered at the sender for retransmission. Figure 2(e) shows the case when TCP loss occurred in Net_m, in which the MP could observe out-of-order IPID. More specifically, the IPID in the TCP packet with Seq_{13} is larger than the IPID in the TCP packet with Seq_{14}. Figure 2(f) shows the case when TCP loss occurred in Net_o, in which the MP could observe duplicated packets in addition to out-of-order on IPID.
- Mass packet loss: Figure 2(h) and Figure 2(g) show the scenarios when multiple consecutive transferred packets were lost. If the packet loss occurred in Net_m (Figure 2(h)), the MP observed abnormal time gap between transferred data. If the packet loss occurred in Net_o (Figure 2(g)), the MP observed duplicated data transfer in addition to abnormal time gap between transferred data.

By no means, we try to list all possible TCP packet loss scenarios. Instead, we would like to point out from these illustrating examples that (1) TCP behaves differently when TCP packet loss occurs in Net_m or in Net_o; and (2) such TCP behavior differences can be characterized via a small set of easy-to-check TCP related network events. We will show in TABLE I that all the scenarios shown in Figure 2 can be identified in our proposed TCP Behavior Model (TBM).

B. Characterizing TCP Behavior

We want to characterize TCP behaviors so that the MP can effectively detect TCP packet loss and identify the corresponding occurring locations based on the observed TCP behaviors.

In the following, we first define several TCP related parameters, and then use them to specify four easy-to-check TCP events that can be used by the *TCPBisector* to detect TCP packet loss and further ascribe the responsibility for TCP packet loss to Net_m or Net_o.

For i^{th} observed TCP packet (denoted as p_i) at the MP, we denote by I_i and Q_i the corresponding IPID and TCP sequence number, respectively. Let $T_{i,j}$ be the time interval between p_i and p_j $(i < j)$ in the same TCP flow, and let T_{f_k} denote the estimated sender's retransmission timeout for TCP flow k.

We introduce four easy-to-check TCP events as below. Each event is denoted by a binary variable $e_i (i = 1, 2, 3, 4)$, and we say e_i is $True$ if the associated network condition is detected. More specifically, we have:

- e_1 (timeout event): When the condition $T_{i,j} > T_{f_k}$ $(p_i, p_j \in f_k)$ is observed, $e_1 = True$.
- e_2 (3-ACK event): When the condition $I_j - I_i \geq 3$ is observed, $e_2 = True$.
- e_3 (reordering event): When the condition $Q_i > Q_j$ is observed, $e_3 = True$.
- e_4 (retransmission event): When the condition $Q_i = Q_j$ is observed, $e_4 = True$.

In our TCP behavior model or TBM, e_1 and e_2 are called triggering events because either event indicates the occurrence of TCP packet loss. e_3 and e_4 are called distinguishing events because e_3 should be observed if the packet lost in Net_m; otherwise e_4 should be observed.

Next, we are going to define two distinguishable TCP behaviors. Each TCP behavior should be observed by the

MP to infer the associated occurring location of TCP packets loss.

Definition 1. We define *Behavior I* as the observable TCP behavior when TCP packets are lost before the MP (i.e., in Net_m), denoted by a binary variable B_I . More specifically, B_I is True when the condition $(e_1 \vee e_2) \wedge e_3$ is satisfied, namely, $B_I = (e_1 \vee e_2) \wedge e_3$.

Definition 2. We define *Behavior II* as the observable TCP behavior when TCP packets are lost after the MP (i.e., in Net_o), denoted by a binary variable B_{II}. More specifically, B_{II} is True when the condition $(e_1 \vee e_2) \wedge e_4$ is satisfied, namely, $B_{II} = (e_1 \vee e_2) \wedge e_4$.

Following up the previously discussed TCP loss scenarios (as shown in Figure 2), now we can characterize them using TBM as shown in TABLE I.

TABLE I: TCP PACKET LOSS SCENARIO IN TBM

TCP packet loss scenario	e_1	e_2	e_3	e_4	B_I	B_{II}
Figure 2(a)						
Figure 2(b)	√			√		√
Figure 2(c)		√	√		√	
Figure 2(d)		√		√		√
Figure 2(e)	√		√		√	
Figure 2(f)	√			√		√
Figure 2(h)	√		√		√	
Figure 2(g)	√			√		√

As we discussed earlier, TCP packet loss can occur in various scenarios. Instead of studying the causality of TCP packet loss, we adopt into our model (i.e., TBM) the essential common events observable in various TCP packet loss scenarios. Thus, the TBM can be used effectively and efficiently identify TCP packet loss, without suffering from the overhead of distinguishing TCP loss scenarios as in many related work. For instance, TCP load balancing, as a misleading scenario discussed in [2], will not trigger any events from $e_1 \sim e_4$ in the TBM if no TCP packet loss occurred.

IV. THE SYSTEM

The *TCPBisector* consists of three modules as shown in Figure 3: (1) data processing module (DPM), (2) inference engine module (IEM), and (3) reporting & querying module (RQM). The *TCPBisector* can be run directly on the MP or installed on a different server. The *TCPBisector* receives from the MP all captured TCP packets, and reports aggregated and flow-based TCP packet loss ratios on Net_m and Net_o, respectively.

DPM collects all TCP packets passing through the MP, and classifies them into TCP flows based on five-tuple (i.e., source and destination IP addresses, source and destination ports, protocol number). The memory location of each recorded TCP flow is stored in a hash table for fast retrieval. For each TCP flow, the *TCPBisector* only needs to keep a fixed number (i.e., 25 as discussed later) of TCP

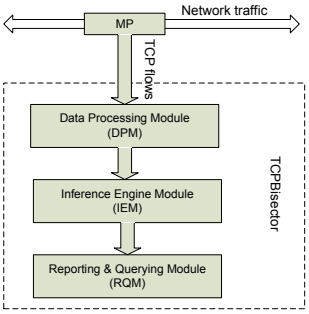


FIGURE 3: *TCPBisector* System

packets in order to accurately estimate TCP packet loss ratios. The corresponding record for each TCP packet in flow k includes its IPID, TCP sequence number, its arrival time, and the estimated retransmission timeout per flow, the TCP packet loss ratios (LR_m^k and LR_o^k). For each flow k, the *TCPBisector* counts the total number of traversed TCP packets denoted as N_k in both directions. We denote by LN_k^I and LN_k^E as the total number of lost TCP packets in Net_m and Net_o, respectively. Accordingly, we have $LR_m^k = LN_k^m/N_k$ and $LR_o^k = LN_k^o/N_k$. The *TCPBisector* also aggregates the flow statistics to provide the aggregated LR_m and LR_o for all observed active TCP flows.

IEM is essentially an event handler. If a triggering event (e_1 or e_2) detected for TCP flow k, IEM verifies the occurrence of distinguishing events (e_3 or e_4) in the recorded flow data structure in order to ascribe the packet loss to the corresponding responsible party (i.e., Net_m or Net_o).

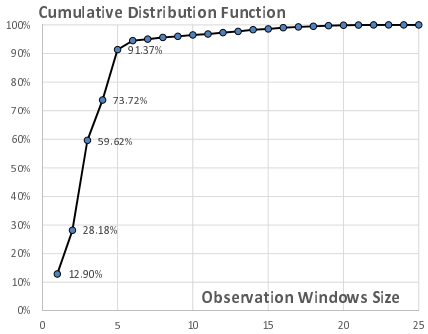


FIGURE 4: CDF of Observation Window Size and Packet Loss Estimation Accuracy

The core function in the *TCPBisector* is TCP loss ratio estimation, which requires to efficiently detect the relevant network events based on captured TCP packets per flow. One significant difference between the *TCPBisector* and the related work [5][15][16][18] is that the *TCPBisector* does not need maintain complete TCP flow information. As shown in our empirical study, the *TCPBisector* only needs to keep track of small number of TCP packets per flow to accurately estimate TCP packet loss ratios. Such a desirable feature in the *TCPBisector* results from the TCP

 ISBN: 978-1-61208-398-8

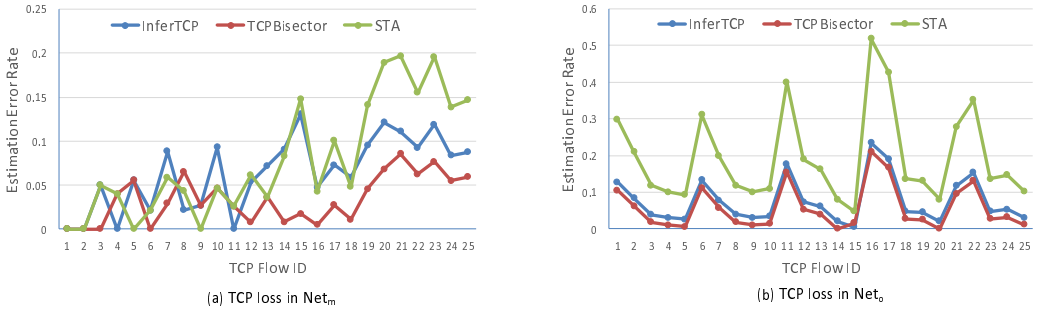


FIGURE 5: Comparison between InferTCP, STA and TCPBisector (a) when TCP loss in Net_m, (b) when TCP loss in Net_o

behavior modeled, which only relies on several easy-to-check TCP events (i.e., $e_1 \sim e_4$). Such a difference makes the *TCPBisector* a lightweight and efficient tool with much lower requirement on the system's computing power and memory space,

Based on the network traces collected from both our emulation and real network experiments, we empirically study the relationship between the estimation accuracy and the size of observation window measured by the number of required TCP packets per flow. The statistics results are shown in Figure 4, where the horizontal axis shows the varying sizes of observation window per TCP flow, and the vertical axis is the CDF of the observation window size measured by trackable number of TCP packets for each flow. Based on Figure 4, we can clearly see that 91.373% of all TCP loss cases, the gap between a lost packet and its retransmitted packet is less than or is equal to 5. We can track almost all TCP packet loss if we record 25 TCP packets per flow. Thus, the time complexity of the inference engine is $O(n)$, where n is the total number of interested TCP flows. Since for each TCP packet, we only keep 20-byte IP header and 20-byte TCP header, the total memory space is bounded by the number of interested TCP flows. For the purpose of cloud application monitoring, the number of flows should be limited.

Finally, RQM updates the per-flow and aggregated statistics of TCP packet loss ratios. RQM also provides query interface such that collaborative parties can correlate their *TCPBisector* reports on the commonly interested TCP flows in order to present a finer-grained view on their network.

V. EVALUATION

We validate the *TCPBisector* using both emulations in a controlled environment and experiments in a Tier-1 network. We also compared the performance of the *TCPBisector* to two related work [15][18].

A. Emulation

We validate the correctness of our methodology used in the *TCPBisector* via a emulation, in which we can obtain the ground truth of various TCP related parameters and packet loss ratios on different network segments.

TABLE II: ACCURACY VERIFICATION VIA EMULATION

LR_m			LR_o		
Actual	**Estimate**	**Error Rate**	**Actual**	**Estimate**	**Error Rate**
0.437%	0.437%	0	0.521%	0.576%	0.106%
0.648%	0.648%	0.000%	5.321%	5.374%	0.010%
0.450%	0.426%	0.053%	8.547%	8.599%	0.006%
0.954%	0.929%	0.026%	5.389%	5.443%	0.010%
1.064%	0.967%	0.091%	8.982%	9.109%	0.014%
3.421%	3.244%	0.052%	1.035%	1.090%	0.053%
3.069%	2.798%	0.088%	4.919%	4.920%	0.0002%
2.778%	2.425%	0.127%	9.590%	9.443%	0.015%
5.631%	5.218%	0.073%	0.817%	0.954%	0.168%
4.931%	4.336%	0.121%	8.207%	8.210%	0.003%
5.413%	4.900%	0.095%	5.307%	5.438%	0.025%
9.071%	8.311%	0.084%	4.884%	5.042%	0.032%
8.819%	8.049%	0.087%	9.249%	9.351%	0.011%
Average Error Rate		0.069%	**Average Error Rate**		0.035%

FIGURE 6: Emulation Environment

In our emulation as shown in Figure 6, a 5-node network is constructed, including two end hosts connected through three routers (i.e., R1, R2, and R3). The TCP packet loss ratios on different router ports are controlled by *Netem* [19]. The error rate is calculated as $Err(LR_m) = |LN^m_{actual} - LN^m_{TCPBisector}|/LN^m_{actual}$ and $Err(LR_o) = |LN^o_{actual} - LN^m_{TCPBisector}|/LN^m_{actual}$ for LR_m and LR_o error rates, respectively. The emulation results as in TABLE II showed that the error on estimating Internal Loss Ratio (LR_m) is 0.069, and the error on estimating External Loss Ratio (LR_o) is 0.035. The result shows that the *TCPBisector* achieves high accuracy on loss ratio estimations for both LR_m and LR_o.

B. Comparison via Emulation

We compare the performance of *TCPBisector* to the two closest related work referred to as *InferTCP* [15] and *STA* [18]. *InferTCP* kept track of the values of two important variables: the senders congestion window (cwnd) and the connection round trip time (RTT) to diagnose end-user-perceived network performance. *STA* [18] developed an

efficient packet classification technique which is used to infer the loss and reorder rates of individual TCP flows.

We adopt the same emulation environment as used in *InferTCP* [15] to compare *InferTCP* and *STA* with *TCPBisector*. We generated 25 TCP flows, and each flow has $3.600 \sim 4,500$ packets. We control the loss ratio is between 0.5% and 10% for each TCP flow. As shown in Figure 5, *TCPBisector* outperformed both *InferTCP* [15] and *STA* [18], and achieved $3 \sim 10\%$ lower estimation error rate on both LR_m and LR_o.

C. CERNET Traces

The traces have been collected at different time from a Tier-1 backbone network CERNET. The MP is placed between the border router in Jiangsu CERNET and the national backbone router. In this paper, we analyze three 5-minute traces collected at properly selected times: 23:55:15, 12, Apr, 2014 (trace 1), 13:55:16, 20, Apr, 2014 (trace 2), 15:55:16, 21, Apr, 2014 (trace 3), representing low, peak, average traffic periods, respectively. The traffic is also classified into forward flows (FF) if destined to Net_o and backward flows (BF) if destined to Net_m.

TABLE III: ACCURACY VERIFICATION VIA EXPERIMENTS

Metrics		Trace 1	Trace 2	Trace 3
# of detected flows		5,504	10,274	9,876
FF	# of Packets	7,872,722	13,441,114	13,437,532
	# of Bytes	4.72 GB	8.14 GB	8.11 GB
	Avg Reordering Ratio	4.012%	3.498%	3.949%
	Avg LR_m	1.686%	1.533%	1.661%
	Avg LR_o	2.705%	2.443%	2.571%
BF	# of Packets	9,650,460	16,584,584	16,578.534
	# of Bytes	7.22 GB	12.79 GB	12.73 GB
	Avg Reordering Ratio	1.494%	2.555%	3.117%
	Avg LR_m	0.836%	1.338%	2.001%
	Avg LR_o	1.220%	1.749%	1.981%

We use the three traces to evaluate the algorithm. The ground truth is hard to obtain in a real network environment with uncontrollable networks. We assume that the performance on the same network remains relatively stable within a short time window (i.e., 15 minutes). Accordingly, we conducted active TCP probing within the 15-minute window after each trace passively collected. Comparing the error between the active and passive measurements for both LR_m and LR_o as shown in TABLE III, the difference is very similar to the results reported in our emulation (5.7% error rate for LR_m and 4.1% for LR_o).

VI. CONCLUSION

In this paper, we propose a lightweight passive monitoring system called *TCPBisector*, in which TCP packet loss responsibility is divided between an internal and external networks conceptually delimited by a network monitor, and quantified using LR_m and LR_o. Using our proposed TCP behavior modeling, the inference algorithm in the *TCPBisector* could accurately and efficiently estimate TCP packet loss ratios with estimation error rate $3.5 \sim 6.9\%$, but only presents computation complexity of $O(n)$ and requires a bounded memory space.

The *TCPBisector* is designed as a coarse-grained TCP performance diagnosis tool. However, *TCPBisector* provides flow based TCP loss ratio estimation. If multiple collaborating parties (e.g., between a cloud user and her service provider) deploy the *TCPBisector* systems, combining the *TCPBisector* reports from both sides on TCP packet loss in a cloud application flows will provide finer-grained view to narrow down the scope of the responsible party.

REFERENCES

[1] J. Sommers, P. Barford, N. Duffield, and A. Ron. "Improving accuracy in end-to-end packet loss measurement." *ACM SIGCOMM 2005*, pages 157-168, 2005

[2] P. Benko and A. Veres. "A Passive Method for Estimating End-to-End TCP Packet Loss." *IEEE Globecom 2002*, pages 2609-2613, 2002

[3] S. Ohta and T. Miyazaki. "Passive packet loss monitoring that employs the hash-based identification technique." *9th IFIP/IEEE International symposium on Integrated Network Management, pages 2-9, 2005*

[4] A. Friedl, S. Ubik, A. Kapravelos, M. Polychronakis, and E. P. Markatos. "Realistic Passive Packet Loss Measurement for High-Speed Networks." *Computer Science*, Volume 5537/2009, pages 1-7, 2009

[5] S. Jaiswal, G. Iannaccone, C. Diot, J. Kurose, and D.Towsley. "Measurement and classification of out-of-sequence packets in Tier-1 IP backbone." *IEEE/ACM Transactions on networking*, Vol.15, NO.1, pages 1199-1209, Feb. 2007

[6] M. Allman, W. M.Eddy, and S. Ostermann. "Estimating Loss Rates With TCP." *ACM Performance Evaluation Review*, pages 12-24, Dec. 2003

[7] H. Nguyen, M. Roughan. "Rigorous statistical analysis of internet loss measurements." *IEEE/ACM Transactions on Networking*, Volume 21 Issue 3, pages 734-745, June 2013

[8] D. Zhang and D. Ionescu "Online Packet Loss Measurement and Estimation for VPN-Based Services." *IEEE Transactions on Instrumentation and Measurement*, pages 2154-2166, Aug. 2010

[9] L. Gharai, C. Perkins, and T. Lehman. "Packet reordering, high speed networks and transport protocol performance." *Proc. IEEE 13th Intl Conf. On Computer Comm, and Networks, ICCCN 2004*, pages 73-78, Oct. 2004

[10] A. Morton, L. Ciavattone, G. Ramachandran, S. Shalunov, and J. Perser. "Packet reordering metrics." *RFC 4737*, 2006

[11] A. Jayasumana, N. Piratla, T. Banka, A. Bare, and R. Whitner. "Improved Packet Reordering Metrics." *RFC5236*, 2008

[12] G. Almes, S. Kalidindi, and M. Zekauskas. "A One-way Packet Loss Metric for IPPM." *RFC2680*, 1999

[13] R Koodli and R Ravikanth. "One-way Loss Pattern Sample Metrics." *RFC3357*, 2002

[14] B. Veral, K. Li, and D. Lowenthal. "New Methods for Passive Estimation of TCP Round-Trip Times." *Passive and Active Network Measurement*, pages 121-134, 2005

[15] S. Jaiswal, G. Iannaccone, C. Diot, J. Kurose, and D. Towsley "Inferring TCP connection characteristics through passive measurements." *In the Proceedings of Infocom*, pages 1582-1592, 2004

[16] M. Ghasemi, T. Benson, and J. Rexford. "Real-time diagnosis of TCP performance in clouds." *In the Proceedings of CoNEXT Student Workshop*, pages 57-58, Dec. 2013

 ISBN: 978-1-61208-398-8

[17] Information Sciences Institute, University of Southern California "Transmission Control Protocol." *RFC793*, 1981

[18] E. Brosh , G. Lubetzky-sharon, and Y. Shavitt "Spatial-temporal analysis of passive TCP measurements." *In the Proceedings of Infocom*, pages 949-959, 2005

[19] A. Jurgelionis, J. Laulajainen, M. Hirvonen, and A. I. Wang. "An Empirical Study of NetEm Network Emulation Functionalities." *In the 2011 Proceedings of ICCCN*, pages 1-6, 2011

 ISBN: 978-1-61208-398-8

Decision-Theoretic Model to Support Autonomic Cloud Computing

Alexandre Augusto Flores, Rafael de Souza Mendes, Gabriel Beims Bräscher, Carlos Becker Westphall, Maria Elena Villareal
Department of Informatics and Statistics
Federal University of Santa Catarina
Florianopolis, Brazil
e-mail: alexandre.flores@posgrad.ufsc.br; rafaeldesouzamendes@gmail.com; brascher@lrg.ufsc.br; westphal@inf.ufsc.br; maria@lrg.ufsc.br

***Abstract*— Much effort has been made to provide a Cloud Computing (CC) autonomic management. Thus, related works are discussed and the need of a full autonomic model with stakeholders is presented. Moreover, this paper introduces a full model of cloud environment to support decision making in autonomic systems. This model is based on an economic utility view of cloud computing, control theory and autonomic computing. It innovates by introducing the concept of conjuncture and imaginary elements (essential elements to forecast and to a non-stationary model). Mathematical modeling is used to formally define a model and a model implementation overview is given.**

Keywords-cloud computing; autonomic computing; decision-theoretic planning; cloud model.

I. INTRODUCTION

The widespread use of computing devices has introduced a drastic change in the way that computing is produced, distributed and consumed. A strong trend is the concept of cloud computing (CC), which is basically a paradigm that deals with economical activity of production, distribution and consumption of computing. According to Kephart et al. [1], the difficulty of managing computer systems goes beyond managing software isolates. The CC dynamic integrates heterogeneous environments and introduces new levels of complexity, outperforming the levels of human capacity [2]. The result is a demand by autonomics clouds.

Although many works propose the automation of CC management, none of them has a model that represents all the stakeholders involved.

This work presents a new CC view based on economy, and utility leading to a useful approach to cloud management. Using a holistic definition, we propose a model to CC management derived from our model introduced in [3]. This generic model can be used to subsidize many decision-making processes and is presented using a mathematical modeling of principal elements and their relationship with eachother.

This paper is organized as follows. Section II addresses the relevant literature and presents our view of CC. Section III presents CC needs for autonomic management based on related works. Section IV describes our proposed model with mathematical representations and presents a simplified class diagram. Finally, we draw conclusions and suggest possibilities for future research.

II. LITERATURE REVIEW

A. Cloud Computing definition

In this section, we will introduce three CC definitions chronologically. Those references brief our view of the evolution of CC definition over the last years.

Foster et al. [4] have an interesting definition for CC: a widely distributed computing paradigm driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet.

Fosters definition is relevant mainly for two reasons: Firstly, he defines CC as a paradigm, and secondly, understands the economic influence at cloud

Furthermore, Buyya et al. [5] have a more complete view which recognizes CC as a paradigm for delivering computing resources as an utility, like gas and water.

Later, the National Institute of Standards and Technology (NIST) [6] defines CC as:

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.

The five essential characteristics stated by NIST are: on demand self-service, broad network access, resource pooling, rapid elasticity and measurable service.

As demonstrated, CC definition has changed in the last years from an economic view to a pragmatic and limited understanding. NIST definition is an attempt to allow comparisons between services. However, they recognize the limitation and state that the service and deployment models defined form a simple taxonomy that is not intended to prescribe or constrain any particular method.

Because we see the CC phenomenon more like Foster et al [4] and Buyya et al. [5], our view of CC is:

the economic activity that focuses on mass production, distribution and consumption of computing. This computing has abstracted logical and physical resources and

 ISBN: 978-1-61208-398-8

prominent commercial frontiers between the stakeholders who produce and consume it.

B. Autonomic Computing

The autonomic computing (AC) concept is based in the human autonomic nervous system that governs our heart rate and body, thus freeing our conscious brain from the burden of dealing with these and many other low-level, yet vital, functions [1]. The overall goal of Autonomic Computing is the creation of self-managing systems; these are proactive, robust, adaptable and easy to use.

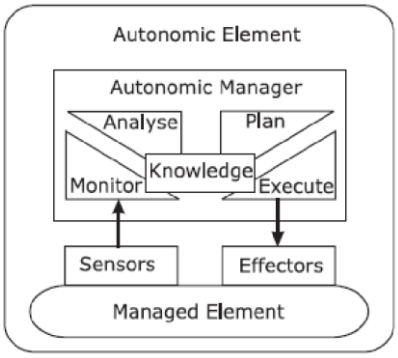


Figure 1. IBM's MAPE-K reference model for autonomic control [6]

A fundamental element that figure in AC bibliography is the MAPE-K control cycle (Figure 1), that consists in Monitor, Analyze, Plan, Execute and Knowledge elements.

For an autonomic system, as shown in [7], to be able to perform self-management, four main abilities must be present: self-configuration, self-optimization, self-protection and self-healing. To achieve these objectives a system must be both self-conscious and environment-conscious, meaning that it must have knowledge of the current state of both itself and its operating environment.

Huebscher et al. [7] define four degrees of autonomicity which can be used to classify autonomic managers and give us the focus, architecturally, that it has been applied. Those elements are:

Support: focuses on one particular aspect or component of architecture to help improve the performance of the complete architecture using autonomicity.

Core: the self-management function involves the core application. It is a full end-to-end solution.

Autonomous: it is also a full end-to-end solution, but the system is more intelligent and it's able to self-adapts to the environment.

Autonomic: this is the most complete level where the interest is in higher-level human based goals like service-level agreements (SLAs), service-level objectives (SLOs) or business goals are taken into account.

C. Control Theory

Control theory uses engineering and mathematics to deal with the behavior of dynamic systems. The objective of a control system is to make de output *y* behave in a desired way by manipulating the plant (system) input *u* [8].

Therefore, we present the first four steps to design a control system, stated by Skogestad [8]:

1. study the system plant to be controlled and obtain initial information about the control objectives;
2. model the system and simplify the model if necessary;
3. analyze the resulting model determine its properties;
4. decide which variables are to be controlled outputs;

Those steps will be mentioned furthermore as the Design Process (DP).

Control Theory often uses transfer functions as a representation, in terms of spatial or temporal frequency, of the relation between the input and output of a linear system. On the other hand, to model complex systems, such as a multi-objetive system, Modern Control Theory often uses a state approach instead of transformation. The system's state is a set of values representing environment.

CC environment management can be classified as a multi-objective multivariable control problem in a time-discrete system. We can assume the dynamics of the CC system to be controlled by several actors where each of the actors has the aim of optimizing its results along the trajectory determined by vectors of control parameters chosen by all players together [9]. A stochastic approach can be used resulting in a Stochastic Multiplayer Game (SMG).

In this class of problem, Nash, Pareto and Stackelberg optimization principles are often used with cooperative and non-cooperative game-theoretic models. To deal with complex systems control, another known strategy is to use Markov Decision Process (MDP) to select the best sequence of actions to been taking. Now we revise those concepts.

1) Nash Equilibrium

Nash equilibrium, proposed by John Nash [10], describes a situation where no player can increase his payoff by unilaterally switching to a different strategy.

2) Pareto optimal

The Pareto optimal is achieved only when a player can become better off in the game without making any other individual worst off.

3) Stackelberg games

A Stackelberg game solution is formulated to model a leader-follower joint optimization problem as a two-level optimization problem between two decision makers.

The upper-level decision maker (leader) announces his decisions to the lower level (follower). Next, follower makes his own decisions and then feeds decisions back to the leader. This implicates in a mathematical program that contains sub-optimization problems as its constraints [11].

4) Markov Decision Process

MDP is a discrete time stochastic control process. MDP provides a mathematical modeling using decision epochs, actions, system states, transitions functions and functions rewards or cost functions.

Broadly speaking, MDP encodes the interaction between an agent and its environment where every action takes the

 ISBN: 978-1-61208-398-8

system to a new state with a certain probability (determined by the transition functions). Choosing an action generates a reward or a cost determinate by reward function.

Policies are prescriptive of which action to take under any circumstance at every future decision epoch. The agent objective is to choose the best sequence of action (policy) under optimum criteria [12].

III. CLOUD COMPUTING CONTROL NEEDS

In this section, we review and show how the scientific community is dealing with autonomic computing to manage Clouds. Firstly, works related to the need of a full autonomic model are presented. Secondly, the need for stakeholders in our model is explained.

A. Full autonomic model

When Sharma [2] designs and implements a system to automate the process of deployment and reconfiguration of the cloud management system, he recognizes that capacity estimation of a distributed systems is a hard challenge. He also states that this challenge is intensified by the fact that software components behave differently in each hardware configuration.

Assuming that we cannot predict how software will perform in any particular hardware, cloud manager be dynamic enough to adapt to these differences. Despite Sharma [2] recognizing this, his approach involves only elasticity performed by nodes allocation based on SLOs, monitoring and forecast.

In [13], autonomic energy-aware mechanisms for self managing changes in the state of resources is developed to satisfy SLAs/SLOs and achieve energy efficiency. Unlike [2], this work focuses on power consumption. It also introduces a more complete model, involving not only physical machines and Virtual Machines (VMs), but expanding on it with customers and a service allocator (interface between the Cloud infrastructure and consumer).

Fitó et al. [14] propose an innovative model of self-management of Cloud environments driven by Business-Level Objectives. The aim is to ensure successful alignment between business and IT systems, extending business-driven IT [15]. In this work, typical IT events and risks during the operation of Cloud providers, such as SLAs or SLOs violations, are not dealt with.

However, Beloglazov [13] shows that many optimization techniques are contradictory. To this end, two techniques are considered: one aimed at the consolidation of VMs and increasing the amount of physical resources in cases of workload peaks; and the other at de-consolidating VMs in cases of node overheating incorporating additional constraints.

Therefore, when the presented models are implemented in ad-hoc approaches, they aim to satisfy only a few users or autonomic computing objectives. As demonstrated, in many cases the models have different granularity levels (hardware level, service levels and business goals). These models cannot be integrated naturally and as a result it is difficult to achieve full management of the Cloud environment.

Palmieri et al. [16] have presented a rich application of game theory to schedule tasks on machines in a multi user environment. They use a temporal model based on time slots to promote each agent interaction scene, but do not consider uncertainty. The game-theoretic approach supports the synergy of agents' objectives in a non-stationary way.

To improve overall system performance, Palmieri et al. [16] introduce a peer-to-peer negotiation method, without a central regulator, that influences agent decisions about its strategies. However, this model is limited by granularity of decisions. Their model is limited because it involves only tasks and schedule.

Thus, we believe that cloud computing needs a full model at the autonomic level as presented by Huebscher et al. [7]. The model is a base for decision-making. A broad, generic, and extensible model can be used with many decision-making processes and can help researchers find the best techniques.

The cloud model must be broad enough to involve all cloud components, stakeholders and their goals. Thereby, it will allow a global understanding permitting the system manager to be able to pay attention to all cloud variables and seek synergy between them. By generic we mean that it must work in any CC system. Extensible characteristic can be understood in two ways: firstly in terms of system variables, the system must deal with undefined variables; and secondly recognising that it is not a final model and specifics scenarios may require new components.

B. Stakeholder

The first step stated in the DP creates the necessity to obtain information about the control objectives. Autonomic computing goals are some control objectives for a CC autonomic manager. Others control objectives are relative and are different in many works, such as [17] [18] [19].

In [20], the following objectives are used for resource allocation and re-provisioning and are represented as use cases:

Acceleration: This use case explores how clouds can be used as accelerators to reduce the application time to conclude by, for example, using cloud resources to exploit an additional level of parallelism.

Conservation: This use case investigates how clouds can be used to conserve allocations, within the appropriate runtime and budget constraints.

Resilience: This use case investigates how clouds can be used to handle the unexpected..

Another example of objectives can be obtained for [13]. A high-level architecture for supporting energy-efficient service allocation in a Green Cloud is proposed. Energy-efficient service allocation is one objective in this work.

Sharma [2] presents two approaches on decisions for dynamic provisioning: cloud provider centric and customer-centric. Cloud provider centric approaches attempt to maximize revenue, while a customer centric approach attempts to minimize the cost of renting servers.

Taking into account Sharma [2], we believe that the objectives presented by Kim et al. [20] and by Beloglazov et al. [13] are relative in what concerns autonomic computing.

 ISBN: 978-1-61208-398-8

This relativism refers to the scope, time and user perspective, or stakeholder.

Stakeholder is a broader concept than users or actors. The term stakeholder involves not only users and cloud consumers, but it also involves the cloud itself, the cloud provider and related parties.

Thereby, we have established the following definition for management of CC as an activity of configuring manageable computational resources to meet and reconcile the interests of various stakeholders, maintaining and increasing the flow of value through the cloud over time.

Thus, we understand that what many authors call objectives, in order to have a complete management at an autonomic level, should be treated as stakeholders' interests.

IV. PROPOSAL

In this section, we present our proposed model and his building process. Aiming to construct a cloud control model that really automates the whole system, we propose a model using as reference the mathematical modeling of Control Theory.

Resulting model of this process is the basis for decision process in CC and it supports the plan phase of MAPE-K. Essential elements of this model are: Stakeholders; Interests; Cloud state; Actions; Events; Conjuncture and Imaginaries elements. Those elements will be presented in the next sections followed by an implementation overview.

A. Essencials elements

1) The Cloud State

The cloud state is a representation of cloud in a specific moment. It represents a static view, just like photography of the Cloud domain. In Markov decision process and in control theory a state is often represented as a tuple of monitored variables and stationary set of all possible states is S. However, in CC, the set of all possible states at time t can be different at the time $t+1$ because monitored variables in a Cloud change in time, creating different sets of possible states.

The controlled variables stated at step 4 of DP are a sub set of monitored variables. Those are represented as dimensions (D_t) in our model. So D_t is the finite set of all monitored variables at time t. For example, (1) represents the resulting set of: CPU of physical machine one ($p1.c$), its memory ($p1.m$) and its state ($p1.s$); CPU of virtual machine one ($v1.c$) and its memory ($v1.m$); and the router usage ($r1.u$).

$$D_t = \{'p1.c','p1.m','p1.s''v1.c','v1.m','r1.u'\} \quad (1)$$

The dimension index X_{dt} represents all possible values of a dimension at time t, where $X_{dt} \in \mathcal{X}_{Dt}$. The relation of D_t and $\mathcal{X}_{Dt}$ is a bijective function ($dx: D_t \rightarrow \mathcal{X}_{Dt}$). So $\mathcal{X}_{Dt}$ can be represented as a set of sets (2), where the first element (X_{1t}) is the index of first dimension at time t, which represents $p1.c$, line 2 is X_{2t} and represents $p1.m$, and so on. This relation is represented by function dx_t.

$$\mathcal{X}_{Dt} = \begin{Bmatrix} '\{'10\%','50\%','80\%','100\%'\}', \\ '\{'20\%',' 40\%',' 60\%',' 80\%',' 99\%'\}', \\ '\{'on','off','reboting','slepping'\}', \\ '\{'10\%','50\%','80\%','100\%'\}', \\ '\{'20\%',' 40\%',' 60\%',' 80\%',' 99\%'\}', \\ '\{'low',' high'\}' \end{Bmatrix} \quad (2)$$

The set of possible states consists of the cartesian product of each set X_{dt} in $\mathcal{X}_{Dt}$. The consequence is that each element s_t in S_t is a tuple $(x_1, x_2, \dots, x_n)$, where x_1 is one element of X_{1t}, x_2 is one element of X_{2t} and so on. Thus, we can represent the S_t as (3).

$$S_t = \prod_{d_t \in D_t}^{d_t} X_{d_t} \in \mathcal{X}_{D_t} \quad (3)$$

2) Stakeholders and Interests

As explained before, ad-hoc objectives are not sufficient to deal with the CC management problem. So, in our model we use a stakeholder interests approach.

The aforementioned acceleration objective, achieved through the allocation of new VMs, is translated in our model as interest of a stakeholder in a state with new VMs. This interest could induce the allocation of more VMs. In this case we can also observe that our model can represent the interests of all involved parties, and the manager could balance the interests using Stackelberg games principles and search for a Pareto optimum or a Nash equilibrium. Allocating more VMs may be interesting for a cloud consumer; however, it can be detrimental to the whole system if, for example, the environment is already overloaded.

Economic problems are normally modeled using a utility function which represents the usefulness of something at a particular time. Extrapolating this concept, we propose an interest function v_t (5) that gives the interest of a stakeholder in a particular state at time t.

$$v_t: U_t \times S_t \rightarrow \gamma \quad (4)$$

As result, (4) returns γ where γ is a real number between -1 and $+1$ ($\gamma \in \mathbb{R} \mid -1 \leq \gamma \leq 1$). So zero represents a neutral interest, positive numbers represent real interest in a particular state and negative numbers represents a non-interest.

We also define a function du_t (5) that maps all dimensions that a stakeholder (u) can change, where U_t is set of all stakeholders at time t and $u \in U_t$.

$$du_t: D_t \rightarrow U_t \quad (5)$$

3) Actions and Events

Once we have introduced the concept of stakeholders, interests and the cloud state, we present the action that allows the connection between these concepts. The stakeholders can affect and change cloud state directly, through actions, and indirectly, through their interests that are passed to the system manager and that can be translated into actions.

 ISBN: 978-1-61208-398-8

Control theory usually chooses a configuration to get the system to a better state. Therefore, MDPs and SMGs usually understand that an action leads to a new state. In [3], we had good adherence to management needs using MDP and actions, but we refined that model and conclude that cloud state can change in an unexpected way because of unpredicted events.

Stakeholders or the system manager can take an action and lead the system to a new desired state with a certain probability, given by function (6).

$$p_t: S_t \times A_t \times S_t \rightarrow \mathbb{R}_{[0,1]} \quad (6)$$

Events are similar to actions and can also change the cloud state. The main difference between them is that events are not planned or even carried out by a stakeholder. An event can be a hardware problem, a software failure or even a power outage, for example.

In addition, the set of all possible actions and events are not stationary, resulting in A_t and E_t. This is because some of them only make sense in some specific state. For example, the action of turning on a server only exists if the server is off at that time. The same occurs with events, a fault in software, for example, can only happen if the software is installed and running. So events and actions are related to states as:

$$s_t Ra_t \subset S_t \times A_t \text{ and} \quad (7)$$
$$s_t Re_t \subset S_t \times E_t. \quad (8)$$

4) Cost fuction

Every action has a related cost. The cost implies in a reduction of a stakeholder interest. Cost function can be defined as (9).

$$c_t: S_t \times A_t \times U_t \rightarrow \mathbb{R} \quad (9)$$

5) Conjuncture and Imaginary Elements(Future)

Here, we define our concept of model conjuncture and its natural derivation, the imaginary elements.

a) Conjuncture

Conjuncture represents the system's structure at a particular time. When new structure elements are added or removed, the conjuncture changes. That is why this element is so important, as what is true in an environment that has, for example, 1 server and 2 VMs, may not be true when the environment grows and has 100 servers and 1000 VMs.

So, for the presented elements we postulate the conjuncture at time t as:

$$cj_t = (D_t, X_t, S_t, U_t, A_t, E_t, dx_t, du_t, v_{t,} c_t, p_t). \quad (10)$$

Other elements can be added to (10) because we are dealing only with essential elements.

b) Imaginary Elements

The following example demonstrates the need of imaginary elements: The environment has one cloud provider and one server. The server at workload peaks uses all available resources and satisfies the SLAs for all consumers. If we give more resources to one of the users, the SLAs will be compromised. The question is: should the system add new resources? Given this, a system manager can infer Nash equilibrium and not allocate more resources. However, a human manager, in that situation, will analyze the whole system, including business goals, and predict new cloud consumers and new demand in the future. So he could identify other needs and have a better plan.

The greatest advantage that a human manager has over autonomic management algorithms is the capacity of human beings to speculate about the future environment. So in order to develop a good plan it is necessary to choose appropriate future actions, based not only on present interests, but possible future interests that may be generated as a consequence of any of these actions.

So, our model can map future imaginary elements, supposing a new conjuncture so that the autonomic manager can take it in to account.

6) Implementation overview

The following implementation overview aims to better explain our model. In Figure 2, a class diagram depicts our proposed model.

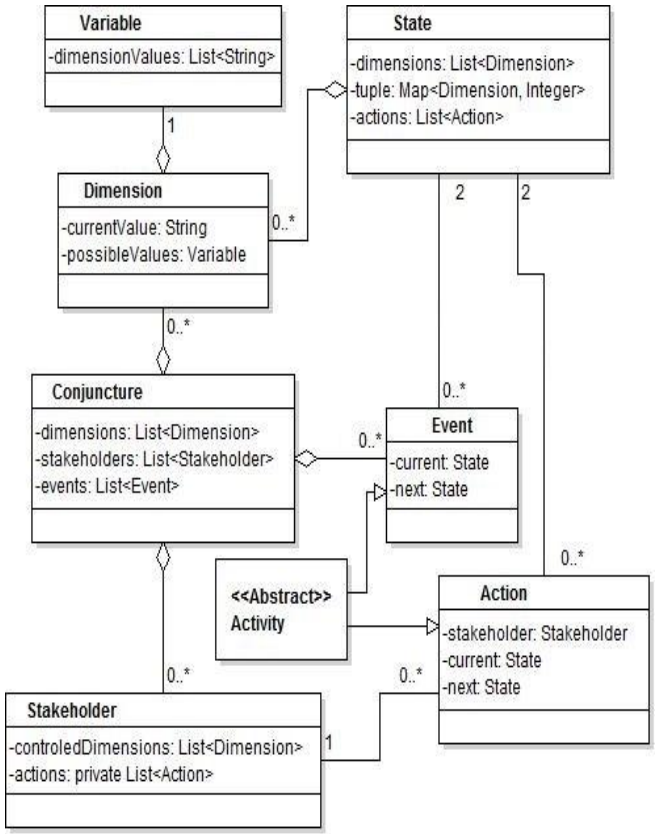


Figure 2. Class Diagram

As shown in (10), conjuncture is the system's core. It has a relationship with dimensions and their possible values, stakeholders, states, actions and events. Although conjuncture class can contain all of them, directly, it is not the only nor the best way to design the system with all elements contained in one class.

 ISBN: 978-1-61208-398-8

Following Figure 2, conjuncture associates directly with events, as they come from an unknown source. Also, it must contain stakeholders, which define a set of controlled dimensions and their actions. Finally, it maps states, indirectly, using all the dimensions from the monitored environment, considering possible states as an aggregate of dimensions. Consequentially, all states can be generated from arranged combinations of possible values in every dimension.

With all sets of components defined, half of the system is modeled. However, the functions, as previously described, by (4), (5), (6) and (9) are not yet defined.

V. CONCLUSION

Based on the view of CC presented, it was possible to base the management model for decision-making on a perspective of public utility management and not only on a data center management perspective.

The presented model gives a solid mathematical base to research political behaviors of CC. Also, using the formalisms that were researched, this work introduced CC management as a multi-player game with high level objectives (Pareto optimal and Nash equilibrium) and presented holistic interests independent of CC architecture or implementation.

Finally, this work presented a new concept of "imagination", essential for a human-like CC management.

For future work CloudSim will be extended to simulate and validate the proposed model and to compare the results with other solutions. CloudSim is a framework to simulation of emerging CC infrastructures and management services.

Following, a multi-strategy approach will be developed. Using Nash equilibrium, Pareto optima, max satisfaction and others in the simulator will be able to choose the best one to implement.

At least, possibilities for future research are:

- Implement a pilot of proposed model using results obtained from simulation;
- Improve the model, if necessary;
- Extend the pilot.

REFERENCES

[1] J. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, pp. 41-50, 2003.

[2] U. Sharma, "Elastic resource management in cloud computing platforms," University of Massachusetts, Ph.D. thesis 2013.

[3] R. Mendes et al., "Decision-Theoretic Planning for Cloud Computing," in *The Thirteenth International Conference on Networks*, 2014, pp. 191-197.

[4] I. Foster, Y. Zhao, I. Raicu, and L. Shiyong, "Cloud Computing and Grid Computing 360-Degree Compared," in *Grid Computing Environments Workshop*, Austin, TX, 2008, pp. 1-10.

[5] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandicc, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599-616, 2009.

[6] P. Mell and T. Grance, "SP 800-145. The NIST Definition of Cloud Computing," National Institute of Standards, Gaithersburg, MD, United States, 2011.

[7] M. C. Huebscher and J. A. McCann, "A survey of autonomic computing - degrees, models, and applications," *ACM Computing Surveys (CSUR)*, vol. 40, no. 3, p. 7, 2008.

[8] S. Skogestad and I. Postlethwaite, *Multivariable feedback control Analysis amd Design*, Wiley New York, Ed., 2007, vol. 2.

[9] D. Lozovanu and S. Pickl, *Optimization and Multiobjective Control of Time-Discrete Systems*.: Springer, 2009.

[10] J. Nash, "Non-cooperative games," in *Annals of mathematics*, 1951, pp. 286-295.

[11] Y. Liu, Y. Ji, and R. J. Jiao, "A Stackelberg Solution to Joint Optimization Problems: A Case Study of Green Design," *Procedia Computer Science*, vol. 16, pp. 333-342, 2013.

[12] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.*: John Wiley & Sons, 2009.

[13] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristicsfor efficient management of data centers for Cloudcomputing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755-768, 2012.

[14] J. O. Fitó, M. Macías, F. Julia, and J. Guitart, "Business-Driven IT Management for Cloud Computing Providers," *CloudCom*, pp. 193-200, 2012.

[15] J. Sauvé, A. Moura, M. Sampaio, J. Jornada, and E Radziuk, "An Introductory Overview and Survey of Business-driven," in *Business-Driven IT Management, 2006. BDIM'06. The First IEEE/IFIP International Workshop on*, 2006, pp. 1-10.

[16] F. Palmieri, L. Buonanno, S. Venticinque, R. Aversa, and B. Di Martino, "A distributed scheduling framework based on selfish autonomous agents for federated cloud environments," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1461-1472, 2013.

[17] G. Wei, A. V. Vasilakos, Y. Zheng, and N. Xiong, "A game-theoretic method of fair resource allocation for cloud computing services," *The Journal of Supercomputing*, vol. 54, no. 2, pp. 252-269, 2010.

[18] A. Corradi, M. Fanelli, and L. Foschini, "VM consolidation: a real case based on openstack cloud," *Future Generation Computer Systems*, pp. 118-127, 2014.

[19] D. Loreti and A. Ciampolini, "Policy for Distributed Self-Organizing Infrastructure Management in Cloud Datacenters," , 2014, pp. 37-43.

[20] H. Kim, Y. El-Khamra, I. Rodero, S. Jha, and M. Parashar, "Autonomic Management of Application Workflow on Hybrid Computing Infrastructure," *Scientific Programming*, vol. 19, no. 2, pp. 75-89, 2011.

[21] Y. O. Yazir et al., "Dynamic resource allocation in computing clouds using distributed multiple criteria decision analysis," in *IEEE 3rd International Conference on*, 2010, pp. 91-98.

 ISBN: 978-1-61208-398-8

Design and Implementation of IoT Sensor Network Architecture with the Concept of Differential Security Level

Jaekeun Lee, Daebeom Jeong, Ji-Seok Han, Seongman Jang, Sanghoon Lee, Keonhee Cho, and Sehyun Park
School of Electrical and Electronics Engineering
Chung-Ang University
Seoul, Republic of Korea
e-mail: li0825, dhmk815, bluephontine, jangsm221, leessan0, thckwall, shpark@cau.ac.kr

***Abstract*— Internet of things sensor networks (ISNs) have been widely used in various fields. Especially, it is a key technology to design environmental monitoring solution in a building space. However, it is not easy to implement ISNs in a building space due to their spatial and structural complexity. Unlike a home space, a building has a variety of variables such as people, rooms, and different structures. Therefore, we considered diverse components of a building which influence 2.4GHz wireless communication and designed efficient ISN structure utilizing sensor node information to provide better network performance. We proposed the ZigBee sensor network system consisting of the environmental information sensor (EIS) and server, designed various user services, and implemented it in a test bed. To verify the efficiency of the system, we conducted two experiments about the network reliability and battery consumption of the EIS, and both results show improvements.**

Keywords- environmental monitoring; IoT sensor network; ZigBee communication; network reliability; differential security level

I. INTRODUCTION

Recently, Internet of things sensor networks (ISNs) have received much attention all over the world. Especially with the development of low power wireless communication and micro electro mechanical systems (MEMS), the ISNs can be configured on a large scale and applied in a variety of areas. ISN is a network in which sensor nodes with computing and wireless communication capabilities are deployed and configure autonomous network. This technology utilizes gathered information from the sensor nodes through wireless communication for monitoring and controlling other devices. Therefore, various services through configuring ISNs such as environmental monitoring, health care service and energy management service are being provided around us [1]. Nowadays, many studies are focusing on the developing environmental monitoring solution in buildings. ISN is key technology to design and develop environmental monitoring solution in buildings due to its ability to manage situational information. However, it is so difficult to implement the ISN in buildings due to their spatial and structural complexity.

A building has complex structural and spatial characteristics compared to a house. Furthermore, in the building, there are various types of people. Thus, there are many challenges to deploy ISNs for the environmental monitoring solution in the building. For example, too many sensors are needed in order to gather and manage the complex environmental and situational events, which inevitably increase costs. Furthermore, the sensor module is more affected by physical characteristics of the wireless link due to structural and spatial characteristics of the building.

In order to manage this complex building information, many researchers studied building information modeling (BIM). BIM [2] is the process of generating and managing building information. BIM encompasses building geometry, spatial relationships, geographic information, and quantities and properties of building components. Geographic information system (GIS) [3] is similar to BIM. GIS gathers the data related to geographic and location-based information. And then it analyzes this information in order to provide user-centric location based service (LBS). It is expected that these systems are evolving into the direction where existing models are applied to various fields such as sensor network management, architecture design, transportation services, etc.

As the number of buildings increases rapidly due to the population growth and industrial factors, there has been a growing interest in safety and energy saving in buildings. Accordingly, the importance of user-centric environmental monitoring services by collecting and analyzing environmental information in buildings is growing bigger and bigger. Therefore, like BIM and GIS, the environmental monitoring service should effectively gather environmental and situational information and provide new services by using gathered information. And this service needs to be operated with energy efficient way, safety of gathered information, and ISN structure suitable for buildings.

In this paper, we considered the design of ISN architecture with the concept of differential security level suitable for the buildings. We proposed the ZigBee-based and reliable sensor network system by utilizing information related to buildings and sensor nodes that configure the network. We utilized the ZigBee technology because of its low-cost and low-power characteristics [4]. Therefore, our system adaptively establishes the network topology, automatically discovers and recovers the faulty nodes according to building information and sensor node information. In this way, we can efficiently manage the ISN and strengthen security of the ISN. And we can also provide better services.

 ISBN: 978-1-61208-398-8

II. PROBLEMS AND REQUIREMENTS OF CONFIGURING ISN IN THE BUILDING

As mentioned above, a building has complex structural and spatial characteristics compared to a house. Thus, there are many problems to configure ISNs for environmental monitoring solution in a building as follows.

- *Interference from other sources*: Radio source transmitting in the same frequency band will interfere with each other. In addition to interference from transmitting source, electromagnetic noise within the building environment can result in interference [5].
- *Multipath propagation*: This phenomenon occurs when portions of the electromagnetic wave reflect off objects and the ground, taking paths of different lengths between a sender and a receiver. This results in the blurring of the received signal at the receiver [5].
- *Faulty node detection and recovery*: Some sensor nodes may fail due to various reasons such as energy depletion, environmental interference, or malicious attacks. This often results in a non-uniform network topology and some nodes will lose contact with the rest of the network. Therefore, the sensor nodes should have a robust and reliable feature to detect faulty nodes and take appropriate measures to recovery from the failure status. This ability is essential to guarantee sensor network reliability and connectivity after one or more nodes are loss in connection with the network [6].

Configuring ISN in consideration of these problems that occur in buildings is essential. Therefore, ISN suitable for buildings needs to have the following requirements*:*

- *Adaptive network management*: Due to the complex structure and spatial characteristics of a building, adaptive network management is essential for securing network reliability. Therefore, in order to configure ISN in a building, various components of a building such as closed or open structure, the number of walls and the number of wireless LAN need to be considered. Furthermore, ISN is required to be managed by considering the status of each sensor node.
- *Energy Efficient Operation*: Extending lifetime of a sensor node in ISN is very important element. Discontinuity of data transmission due to the lack of battery reduces network reliability and can cause incorrect data transmission to users. Thus, battery condition of each node should be analyzed for configuring ISN, and data transmission path needs to be determined based on the battery conditions of surrounding nodes.

III. WIRELESS ENVIRONMENTAL IOT SENSOR NETWORK SYSTEM ARCHITECTURE

The architecture of the proposed system is shown in Figure 1. Environmental information sensor (EIS) consists of 6 type of sensors, such as temperature, humidity, motion sensor, carbon monoxide (CO), and illumination.

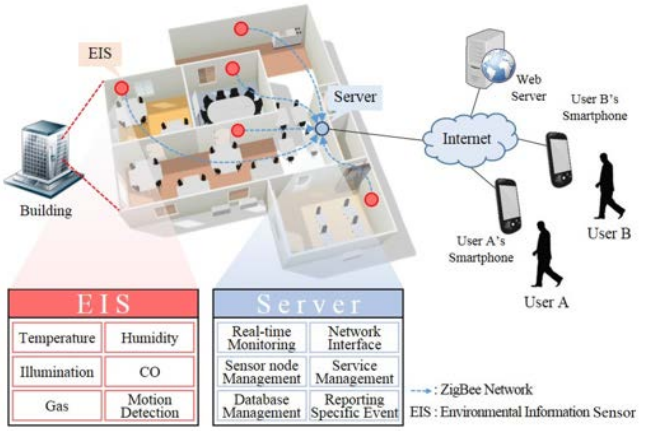


Figure 1. Overview of wireless environmental IoT sensor network system.

Moreover, it contains LEDs and a buzzer for notifying operation of current status. If every environmental sensor is included in the sensor node, it would be wasteful because some environmental sensors could not be used in some spots. Therefore, each sensor is designed to be detachable and also shut out power source in case of that some sensors are not used. Both EIS and server use ZigBee wireless communication for efficient energy use. The EIS plays a role of gathering environmental information about situations that occur in the buildings. The server analyzes and stores the information received from the EISs to provide user services. Users are able to confirm the analyzed data through smartphone application and web server.

A. Network Structure of the Proposed System

The whole network structure of the proposed system is optimized to consider various variables in building spaces.

1) Network Initialization of the Proposed System

A coordinator manages the certain number of sensor nodes or nodes in a particular space. However, in this structure of sensor networks, if a coordinator does not work, it influences the network performance which the coordinator manages. Therefore, each sensor node has the same middleware and hardware specification so that every node can play a role of the coordinator. Figure 2 shows network initialization of the proposed system. At the beginning of the network initialization, a coordinator is selected by the server, and it can be changed according to various cases. The selected coordinator gathers and stores the data from assigned nodes, and transmits to the server.

First, the server is installed, and each sensor node is distributed in a specific area. In this paper, the server uses connectivity between the sensor network and building area so 16 area codes that help to understand where it is deployed have to be selected in the nodes. By using a switch in the EIS node, users can change or choose one of the area codes, and the selected code is added to the event message of the coordinator. Therefore, the server can link the node position with a building floor plan.

 ISBN: 978-1-61208-398-8

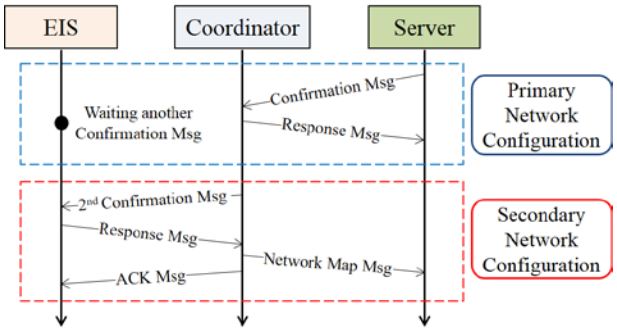


Figure 2. Network initialization of the proposed system.

For the network initialization, the server needs to select a coordinator among the nodes, and this process is as below. First, the server transmits the Confirmation Msg around itself, and the nodes receiving this Msg also send the Response Msg to the server. The selected nodes are candidates of the reserved coordinator. Secondly, these candidates transmit 2nd Confirmation Msg, and other nodes, which did not receive anything in the primary network configuration, send Response Msg to the candidates. The server repeats this process until every sensor node is found, and every node has graded level to be a coordinator. That is, the node connected during primary network configuration has the highest graded level. As described above, if multiple candidates of the reserved coordinator are selected, the server chooses one of them as a coordinator that is located in the largest area and has a wide coverage range, and other candidates gain a qualification of the reserved coordinator. Therefore, if the selected coordinator does not work, the server selects one of the candidates to maintain a certain network. Moreover, one of the EIS nodes, which are in the same section and have same area code, is selected as a sub-coordinator according to the coverage area, and this sub-coordinator collects and transmits data to the coordinator.

2) Building Elements Considered for Configuring Wireless Environmental IoT Sensor Network

For efficient wireless environmental IoT sensor network suitable for buildings, we considered several components of a building. First of all, we found factors which have effects on 2.4 GHz frequency in a building area. High frequency communications such as ZigBee and WLAN are influenced by various factors more compared with low frequency communications. Furthermore, positions where sensor nodes are placed would be important in network performance. Therefore, we analyzed researches related to ZigBee communication and chose some factors [7], and designed the proposed sensor network according to the factors. The types of evaluation factors are as follows.

- Coverage area: grasping the optimal number of nodes and service quality
- Closed or open structure: if a node is placed in a closed structure, the server least chooses the node as a coordinator
- Wall quality and the number of walls: if a node is placed near aluminum quality walls, the server least chooses the node as a coordinator
- The number of wireless LAN used in a space: related to communication interference in the same frequency band
- Degree of communication interference between floors: related to communication interference in different floors

3) Hierarchical Network Structure

In this paper, the wireless environmental IoT sensor network system is hierarchized based on the status of sensor nodes such as battery status, the number of performed events and node area. Figure 3 shows the hierarchical network composition. The green node means a coordinator, and colored nodes means hierarchical nodes according to the network initialization. For example, the blue node which is in the Alternative path is included in two common paths, and the server assigns this node in the one of the two networks, which has lower battery status. If a node placed near the Alternative path does not work, the coordinator replaces the node with one in the Alternative path. If there is no node in the Alternative path, the server reports that a node needs to be changed. And, each node has a differential security level. The coordinator has a high security level because it deals with a lot of information. Furthermore, in order to improve battery life and reliable data transmission, environmental sensors of the EIS nodes are controlled by the coordinator. If battery status of an EIS node is below 20%, the node turns off the environmental sensors according to the control command of the coordinator. By doing this, the node, which can be in a critical path, can save power and focus on data transmission instead of wasting power for sensing environmental information.

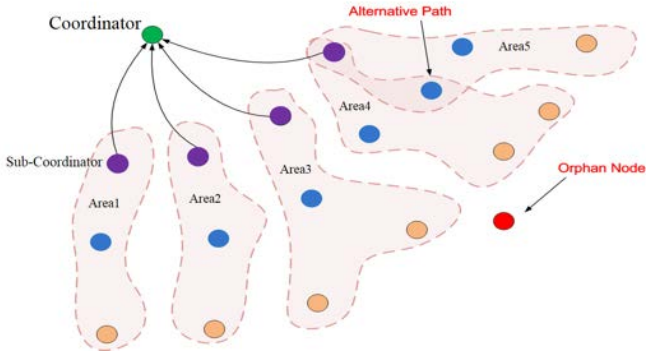


Figure 3. Area-code based hierarchical network structure.

In organizing the sensor network, the orphan problem is always an important issue. Especially, if a node is in a critical route path and has a problem, a network which is connected to the node cannot operate well. In a building space, it is not usual to place hundreds of sensor nodes at the same place so that the orphan problem is getting more important. In some cases like the red node, some nodes could not reply in the first network initialization and be included in the network, and this causes another type of the orphan problem. However, we designed this sensor network system by grouping the certain number of nodes in the same place like Figure 3, and users have to input the total number of nodes to compare the initial number of the nodes with the

 ISBN: 978-1-61208-398-8

number of the installed ones. Thus, the server can detect how many nodes are not included in the network.

B. Data Packet Structure of the Proposed system

Figure 4 shows the data packet structure of the proposed system and management of the data packets in the server. The data packet of an EIS includes area code, data length, event code, environmental information and battery status. The coordinator creates data packets by gathering information from EISs and send them to the server. Since the coordinator does not have sufficient internal storage space and processing performance, all the data is translated into the hex codes and deleted after the data transmission. As described above, each node can select one of the 16 area codes by a switch and every area is also allocated for one of 16 area codes. After the network initialization, the server checks whether data is received or not through coordinators. That is, each coordinator checks and stores detailed network connections between a node and the coordinator. Therefore, these data are transmitted to the server, and the server figures out what events happen in each node and can analyze network statistics. Furthermore, the coordinator counts time by using internal timer so that time can be added to each event.

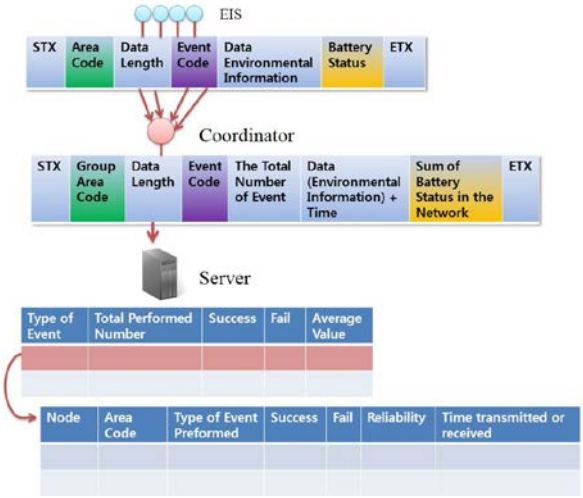


Figure 4. Data packet structure of the proposed system.

Basically, the server not only gathers environmental information but also analyzes the data packet so that every event can be inferred. For example, the EIS sends a temperature data periodically but motion-detection or gas data is transmitted when specific events are discovered. Furthermore, the coordinator also gathers a battery status of each node because a Micro Control Unit (MCU) of each node checks and sends the battery status. That is, if a specific node takes a role of a router and runs down battery, the server reports this data to replace the battery or change power source.

The server also can find communication problems in specific routes. For example, if the server knows the total number of temperature events in a node, it can also infer the ratio of data transmission success or fail. Furthermore, the server can find a period of each event based on the time counted by the coordinator. Therefore, the server arranges every event in time table so that it is easy to figure out what events happen at the same time and change a time schedule of events that have a problem.

IV. Service Provisions of the Proposed System

In this paper, we also designed various user services through the smart phone application and web service as shown in Figure 5. Especially, smart phone is widely used in these days to provide mobile services and various applications [8]. Furthermore, users do not need to be in limited places to access web sites, and it is available to confirm what they want to check in various places and let users know certain events by using push service.

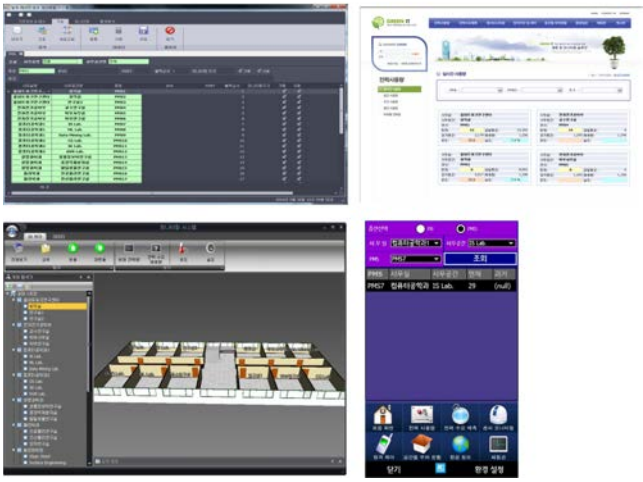


Figure 5. Management program, web site, and smartphone application

The services that the server and EIS provide are as follows.

- Provision of collected environmental information
- Sensor node management
- Reporting specific events

As we described above, total 6 environmental sensors are included in the EIS, and the system provides real time monitoring services, such as motion detection of users, and risk factors like gas and fire. According to the purpose and importance, the environmental information is divided into three parts, and each part is managed differently.

- General environmental information: Temperature, humidity, illumination
  - Indirectly used in other application
  - Periodically gathered
  - Low data grade
- Event-based environmental information: Motion detection
  - Detected when specific events happen
  - Indirectly or directly used in other application (crime prevention, people density)
  - Information period is not irregular
- High priority environmental information: CO, gas
  - Detected when specific events happen
  - Indirectly or directly used in other application

 ISBN: 978-1-61208-398-8

- reported immediately to users or the server
- High data grade

The general environmental information is not used immediately but first gathered periodically to help to manage battery status in the EIS. This information is gathered by the coordinator, and sent to the server to provide real time environmental monitoring or used in additional services.

The event-based environmental information includes only the motion detection sensor, and it is gathered irregularly and used indirectly or directly according to applications. Therefore, according to services, the coordinator sends this information instantly or periodically to the server.

The high priority environmental information consists of two sensors, CO and gas. If a fire breaks out, CO is generated so that the server can detect fire by using a CO sensor. Therefore, CO and gas information have high priority because these are connected with risk factors in a building space. If some nodes are performing other events, all events are stopped, and the coordinator transmits this high priority information instantly to the server.

V. IMPLEMENTATION

First of all, figure 6 shows the hardware structure of the EIS. The EIS includes,

- MCU: controlling each part of the EIS
- Power Part: consisting of battery and power source
- Battery status part: checking the battery status
- Sensor part: including 6 types of environmental sensors and converting analog input to digital
- ZigBee part : ZigBee communication

The server is designed by using C++ based programming language, and a ZigBee module is attached to communicate with the EIS nodes.

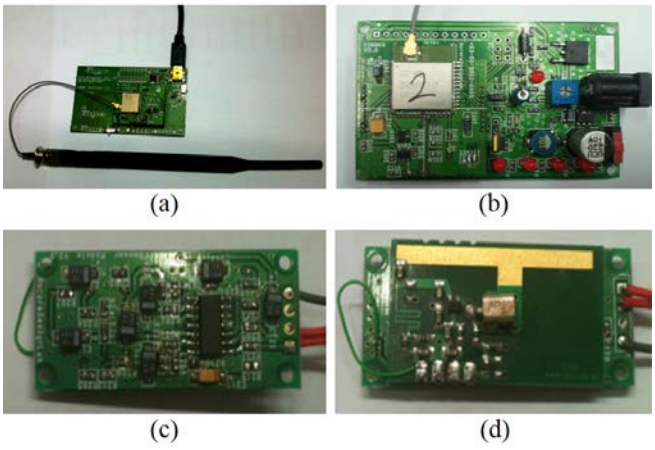


Figure 6. (a) ZigBee module with an antenna used in the server, (b) Hardware structure of an EIS, (c), and (d) Hardware structure of sensor part.

The EIS nodes and server are implemented in the test bed environment. Figure 7 shows the floor plan of the test bed. Through the network initialization, total 8 coordinators are selected in 8 areas, and the server is located in the red field. In each section, 10 EIS nodes with 6 types of environmental sensors are placed. Above this, the experimental environments are in Table 1.

TABLE I. EXPERIMENTAL ENVIRONMENT

Classification	
The Number of Used WLAN	1~3
Status of Wall Quality	Normal
Extent of Testbed	3200 m^2

The component about communication interference between floors is excluded because the test is performed in only one-story house. In summary, there are the total 8 sections and 80 EIS nodes with 6 types of environmental sensors, and we tested to monitor environmental information and control sensor nodes by using the server, web page, and smart phone application.

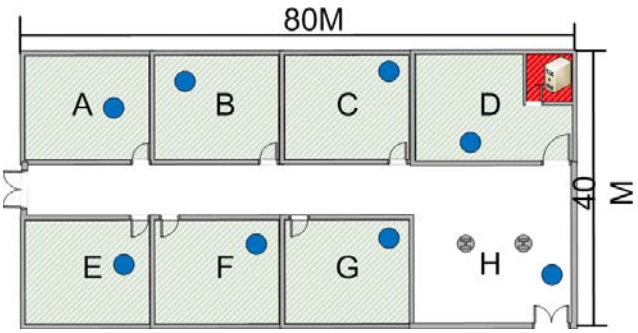


Figure 7. Testbed floor plan; blue points means coordinator, and the server is located in a red section.

VI. TEST AND RESULTS

Based on this test bed environment, we tested 2 experiments about reliability and efficiency of this system. Performed experiments are as follows:

- Network reliability of the proposed network structure
- Comparison of battery consumption

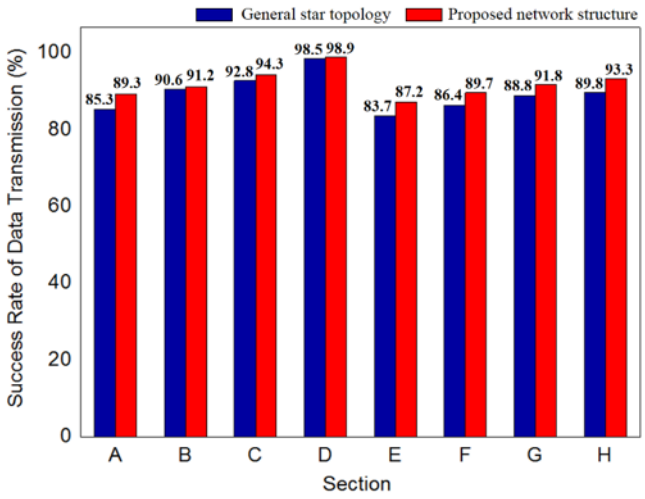


Figure 8. Comparison of data reliablility between two networks.

The first experiment describes how high the system provides network reliability. We compared a network, which uses general star topology, with the proposed network structure. Figure 8 shows the success rate of data transmission between each coordinator and the server. For this experiment, each coordinator transmitted 1000 data packets to the server, and we analyzed this result. The result shows that the highest rate is obtained the D section, and the lowest rate is obtained the E section in Table 2. This result means the number of walls and distances between the coordinator and the server influence the network performance, and the system improves it.

TABLE II. FIRST EXPERIMENTAL RESULTS ACCORDING TO DISTANCE AND THE NUMBER OF WALLS

Section	Distance	The Number of Walls	Success Rate (%)	
			General Star Topology	Proposed Network Structure
A	75M	4	85.3	89.3
B	55M	3	90.6	91.2
C	35M	2	92.8	94.3
D	15M	1	98.5	98.9
E	82M	5	83.7	87.2
F	60M	4	86.4	89.7
G	49M	3	88.8	91.8
H	35M	2	89.8	93.3

The second experiment shows how much the system improves battery life. The EIS node has a 900mAh battery, and the network operated for 1 day. To result this experiment, we compared the sum of battery amount of the two network cases described in the first experiment. Figure 9 shows the comparison of battery amount of two network cases.

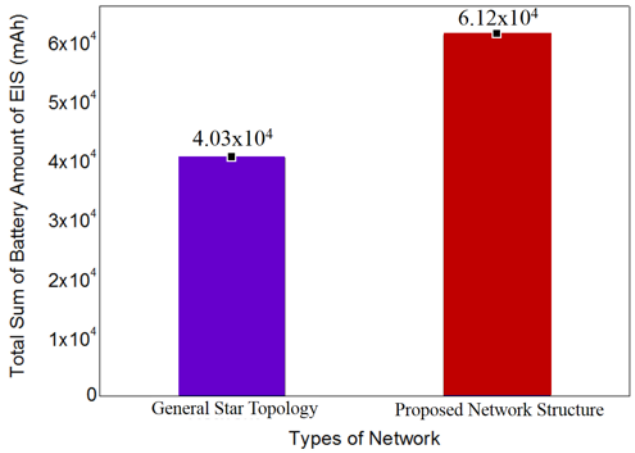


Figure 9. Comparison of battery amount.

This result shows that the proposed network structure saves about 34 % battery life. That is, the server checks service quality of each section and control power of used environmental sensors. Therefore, the proposed network can maintain the entire nodes' battery life longer.

VII. CONCLUSION

In this paper, we designed the wireless environmental IoT sensor network system by using the EIS and server in a building space. The main point is that this sensor network system considered various building elements which influence ZigBee communication and used sensor node information for providing better performance of the network. Users can confirm various environmental conditions such as temperature, humidity, illumination, CO, gas, and motion-detection through the smart phone application or web site. We implemented this sensor network system in the test bed and performed two experiments about performance of the system. The experimental results demonstrate the improved network reliability and longer battery life by using the proposed sensor network system.

ACKNOWLEDGMENT

This work was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1044) supervised by the NIPA(National IT Industry Promotion Agency), and by the Human Resources Development (No.20124030200060) and the Energy Efficiency & Resources (No.20132010101850) of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government Ministry of Trade, Industry and Energy.

REFERENCES

[1] M. Erol-Kantarci and H. T. Mouftah, "Wireless Sensor Networks for Cost-Efficient Residential Energy Management in the Smart Grid," Smart Grid, IEEE Transactions on, vol. 2, pp. 314-325, June 2011.

[2] S. Azhar, M. Hein, and B. Sketo, "Building Information Modeling (BIM): Benefits, Risks and Challenges," In Proceedings of the 44th ASC Annual Conference USA, vol. 2, pp. 2-5, April 2008

[3] C. P. Lo and W. Yeung, Concepts and Techniques of Geographic Information Systems (Ph Series in Geographic Information Science): Prentice-Hall, 2006.

[4] A. Wheeler, "Commercial applications of wireless sensor networks using ZigBee," Communications Magazine, IEEE, vol. 45, no. 4 , pp. 70-77, Apr 2007.

[5] J. Kurose and K. Ross, Computer Networking: A Top-Down Approach: Pearson Education, 2008.

[6] M. S. Pan, C. H. Tsai, and Y. C. Tseng, "The orphan problem in zigbee wireless networks,". IEEE Transactions on Mobile Computing, pp. 1573-1584, Nov 2009.

[7] L. Chan-Ping, J. L. Volakis, K. Sertel, R. W. Kindt, and A. Anastasopoulos, "Indoor propagation models based on rigorous methods for site-specific multipath environments," Antennas and Propagation, IEEE Transactions on, vol. 54, no. 6, pp. 1718-1725, June 2006.

[8] P. Zheng and L. M. Ni. "Spotlight: the rise of the smart phone," Disributed Systems, Online, IEEE, vol. 7, no. 3, Mar 2006, doi:10.1109/MDSO.2006.22.

 ISBN: 978-1-61208-398-8

A Flexible Self-Aligning Communication Solution for Multinational Large Scale Disaster Operations

Peter Dorfinger, Ferdinand von Tüllenburg, Georg Panholzer, Thomas Pfeiffenberger
Advanced Networking Center
Salzburg Research Forschungsgesellschaft mbH
Salzburg, Austria
email: {firstname.lastname}@salzburgresearch.at

***Abstract*—This paper presents a communication solution for large scale multinational and multi-organizational disaster operations. The work is motivated by real world requirements depicted from well experienced forces in disaster management. The core design principles for the solution are flexibility and easy installation. The solution has proven its applicability during several training and large scale exercises, such as floods or earthquakes. With our solution we present a communication infrastructure for connecting mobile devices of relief forces in large-scale disaster operations to Command, Coordinate and Control Systems. The infrastructure can be set up by one single, non-IT-expert person.**

***Keywords*—Emergency network; self-alignment; interoperability; communication; wireless coverage simulation.**

I. INTRODUCTION

Communication is one main building block to enhance interoperability between multinational and multi-organizational disaster relief actions. To provide targeted and fast help, relief organizations are highly reliant on the possibility to share important information across different organizations and national borders.

Compared to former days, when mostly voice radio communication was used for information interchange, modern broadband communication technologies can now provide a clear added value to the cooperation in large scale disaster actions. For example, some of the most valuable disaster information is gathered by relief forces directly at incident locations. Sharing these information between relief workers and management could bring a clear benefit to find the best supporting measures shortly. To get out the best, all the information should be made available in an automated way to all relief organizations and personnel involved in the disaster response actions.

With broadband communication, automatic exchange of relevant data between relief organizations as well as voice communication will be possible. One main research issue is now, how to bring broadband connectivity to almost any arbitrary location in a large scale disaster area.

The presented communication solution is part of the European FP7 research project IDIRA (Interoperability of data and procedures in large-scale multinational disaster response actions) [1]. IDIRA has its overall objective target in enhancing and streamlining the cooperation between relief organizations by enabling interoperability of information systems used for disaster management.

IDIRA addresses this interoperability topic twofold. First, at an organizational level, IDIRA shall examine possibilities to reach administrative coordination of multi-national disaster relief organizations, with all their own specific workflows and procedures. Second, on technical side, IDIRA shall provide a complete solution consisting of information systems, communication protocols, software interfaces, and standard data formats. This solution is the enabler technology to exchange disaster related information between administrative operators, executive personnel, and other disaster management systems connected to IDIRA. With IDIRA, information on incidents, resources, observations, and sensor data should be collected and shared to various other information systems like, mobile devices and command and control systems (C&C). To reach the required level of interoperability and automatic information exchange, IDIRA has a strong focus on a flexible, reliable, and easy to deploy communication infrastructure.

One of the major problems, we address with our proposed communication infrastructure is, that after a large scale disaster, the existing public broadband network is often partially destroyed, overloaded, or at least hit by power outages. Consequently, first responders cannot rely on any pre-existing infrastructure which may fail as consequence of the disaster.

As the communication network is essential for a better and more efficient collaboration between first responders, there is a critical need for the fast setup of alternative communication means. In such a case, another issue arises: First responder organizations neither are experts in setting up communication equipment, nor there are a large number of IT experts available within their field staff. Thus, easy setup and maintenance is heavily required for such systems.

This paper is structured as follows: Section II gives a brief overview on related work, Section III describes the system requirements on the IDIRA communication network, Section IV presents our proposed solution; the results of the systems' usage is discussed in Section V. Section VI concludes the paper and outlines further steps and ideas to improve our solution.

 ISBN: 978-1-61208-398-8

II. RELATED WORK

Communication technologies used by relief organizations is manifold. Beside everyday mobile communication technology like 2G, 3G, 4G, or even standard voice radio, there are numerous technologies which are more specific to action forces or disaster relief organizations. Some of these technologies are designed to be transportable and independent of any pre-existing infrastructure like satellite communication. Satellite communication systems like BGAN [2], VSAT [3], or Emergency.lu [4] are specially designed to provide data and voice communication in remote areas. As such, it can be used as communication uplink in large-scale disasters, if the pre-existing infrastructure is damaged or not usable due to power outages. Drawback of most satellite communication technologies are the high expenses for data exchange, so making them not the number one solution for commanders in the field, but a feasible approach for one Internet uplink in the operational area. BGAN additionally has only a very limited bandwidth.

In the case of a disaster, affecting many people and large areas, the public landline and mobile phone networks are often affected by overload, power blackouts, and damaged infrastructure. Consequently, these networks are often unusable as reliable communication infrastructure. Among others, this problem was addressed by TETRA. TETRA allows both, range limited direct device to device communication without usage of a fixed infrastructure and range unlimited indirect communication via a fixed infrastructure. To be more protected against power outages and damages, the components of the fixed infrastructure are constructed on a redundant basis.

A disadvantage of these technologies is the provided low bandwidth for data exchange. IDIRA heavily depends on data exchange between multiple components - for example for user interactions via IDIRAs web interface, the so called Common Operational Picture (COP). Here, data are exchanged between web clients of tactical personnel at the command & control center and field commanders. The bandwidth provided by TETRA will not be sufficient to operate several end devices in parallel.

Other available communication technologies have drawbacks regarding operating licenses. E.g., licenses are needed for operating WiMAX [5] communication equipment. Highly Mobile Network Node (HiMoNN) [6] is a communication system specific to public protection and disaster relief (PPDR). In compliance with ECC Recommendation (08)04 [7], HiMoNN operates with transmission power of 8W in the 5GHz frequency band, and is able to transmit data over a distance of several kilometers with a bandwidth of 28Mbit/s. Shortcoming of the HiMoNN technology is the lack of international operating permissions, which make it not the best choice for an international deployment.

The authors of [15] propose to use end-user devices such as mobile phones to establish a mobile ad-hoc network (MANET) between first responders. The devices use their 802.11 wireless network interfaces, to automatically build up connections to other devices within their transceiver range. During operation, first responders sending data via this MANET to a central host located in a command center. Special routing protocols are employed for routing the network traffic. To have a practical solution, a rather dense concentration of devices is needed, so that MANETs are only usable at limited incident areas of less than some 100 square meters. As a general communication solution, this approach seems not to be sufficient, as it cannot be assumed to have an adequate density of devices in the field to span a network across all devices.

The solution proposed in this paper uses 802.11 [8] technologies, which can be used all over the world without special licenses, but the possible distance between two devices is more limited than with WiMAX or HiMoNN. As routing protocol we use the Optimized Link State Routing Protocol (OLSR) [9], which is optimized for constrained wireless LANs. OLSR is based on multipoint relays which reduce the routing overhead on the network.

Within IDIRA, disaster information is represented in a standardized and open XML-based messaging format known as Emergency Data Exchange Language (EDXL) [10]. Out of this suite of standards, the EDXL-CAP (Common Altering Protocol) [11] data format is applied to data concerning occurred incidents. These incidents are registered e.g., by a sensor system and shared with some central C&C system. Information respective to availability, demand, and status of resources, such as specialized rescue units or even power generators, is shared by the EDXL-RM (Resource Messaging) [12] standard. The EDXL-SitRep (Situation Report) [13] messaging standard is used within the IDIRA context for exchanging information on observations and situation reports sent by commanders in the field via their mobile devices.

III. REQUIREMENTS

The proposed networking solution is intended for multinational and multi-organizational large scale disaster operations. The work is part of the IDIRA project. Consequently, it has to fulfill both the generic requirements brought in by first responder organizations and the needs pushed by the IDIRA framework on the communication solution.

Within the IDIRA project several end user organizations are involved and more than 20 organizations are part of the end user advisory board. The requirements concerning the communication solution for first responders in emergency situations were conducted within IDIRAs' end user advisory board.

As nobody knows where the next disaster will strike the communication solution has to be allowed for usage (almost) all over the world. This allowance has to be given in advance, as asking for e.g., WiMAX frequencies in a disaster area right after the disaster strikes, is not an applicable approach (REQ1).

Whenever open broadband communication networks are at least partially operational, the emergency network should be able to use or interact with the existing network (REQ2).

For end device connectivity, the network should use open standards to ensure that different end devices can easily be

 ISBN: 978-1-61208-398-8

added to the network and can interchange information with the emergency network (REQ3).

After a large scale disaster, only few human resources are available to setup a communication network. Especially IT-staff is a scarce resource. Thus, the setup of the communication network, especially of the remote nodes, has to be easy. There must not be the need for an IT guy to setup a remote communication node (REQ4).

Even in case the network is down, the information, that was already in the system when the connection was lost, should be available to the end-user (REQ5).

The interaction with the system should be the same when the network is down, as it would be if the network is up - for sure with a limited number of functionalities (REQ6).

The result of a survey within the end user advisory board was that REQ1, REQ2 and REQ4 are the most important ones and are thus crucial for the applicability and acceptance of a proposed solution.

The main design principle of IDIRA is to use standardized interfaces. Standardized interfaces, as for example the XML based EDXL standards family, often have the drawback to increase the communication overhead. Consequently the bandwidth needs within IDIRA are higher than they could be in case the design would have been performed with the bandwidth as main scarce resource. The IDIRA communication network will be used to access the Common Operational Picture (COP). The COP is a web application which presents the needed information in a Geographic Information system (GIS) manner to tactical and strategic personnel. For bootstrapping a device, COP needs about 10MByte of data as initial load. During operation data containing sensor information, incident information, and information on resources and their activities are exchanged. All this leads to a bandwidth requirement of around 2Mbit/s for a seamless operation of the COP. For field commanders (operational commanders) a dedicated Android App has been developed, where specific attention was given to reduced bandwidth consumption. This operational app needs less than 100kBit/s to be operational, if voice communication is used the needed bandwidth is in the range of 200kBit/s. Depending on the number of clients a few Mbit/s of bandwidth are required to support the IDIRA needs (REQ7).

IV. SOLUTION

Core of our proposed communication solution is a set of multiple communication nodes, called Wireless Gateway (WGW). The WGWs are used for both, to build up a backbone network and to provide end users access to this network.

Based on the requirements presented in section III, 802.11 based technology was chosen as the main communication solution used in the WGWs. The usage of 802.11 ensures that it can be used almost all over the world without the need to apply for licenses (REQ1). Within the proposed solution for the backbone network the 5GHz frequency range is used. 802.11 technology also is used between WGW and end devices. We are using the 2.4GHz frequency band to connect to end devices. This technology is widely used in end devices such as smart phones, netbooks, tablet devices or laptops. This ensures that a large number of commodity hardware will be available to be used as end devices (REQ3).

To overcome the limitations of the limited distance between two 802.11 endpoints we are using directional antennas. The disadvantage of directional antennas is that they have to be aligned to each other. As this is not a task which can easily be performed by first responders, it was decided to develop a self-alignment algorithm. The proposed solution consists of three directional antennas at each WGW which are automatically aligning themselves to build up a meshed backbone network with a maximum number of three direct links to other WGWs.

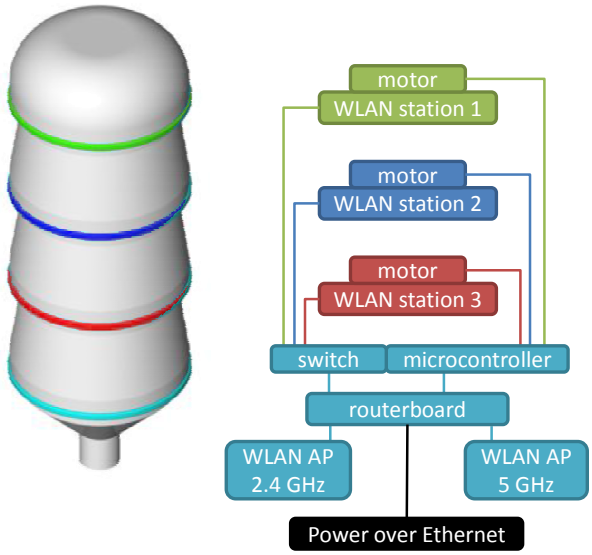


Figure 1. WGW building blocks.

Figure 1 shows the main building blocks of the WGW. The upper three layers are built identically. They consist of a WLAN station with a 16dBi directional antenna and a motor to rotate the antenna. In the bottom layer, a switch is mounted to connect the WLAN stations to a router-board. On the router-board the self-alignment algorithm is running, which controls to which remote WGW the individual WLAN stations are connected. A microcontroller is connected to the router-board which is responsible for performing the rotation of the upper three layers. In addition, the router-board is equipped with two miniPCI WLAN cards, where the first one is configured for the 2.4GHz band, and used for the connection to the end devices, the second operates within the 5GHz band and is used for the self-alignment algorithm. Details on the self-alignment algorithm can be found in [14]. The WGW will span a meshed network automatically, and each WGW is working as relay node for other WGWs.

The installation of the system is quite easy, it is just needed to mount the WGW on top of a pole and connect the cable and the Omni antennas for the 2.4GHz and 5GHz access points. Thus, the installation can be performed by non-IT experts, bringing us closer to REQ4. Nevertheless, an easy to setup network is only half of the job. In advance, it has to be decided where to set up the communication nodes.

 ISBN: 978-1-61208-398-8

To support the first responders in positioning of the WGW, a simulation environment has been developed. The COP visualizes incidents, resources, tasks and other relevant information for disaster management on a map. Based on this information the incident areas are known, where field commanders need a working communication infrastructure to interact with the IDIRA system. Within one single incident area there may be preferred positons for WGWs (where they are able to communicate to each other) and positions where a WGW should not be placed. For first responders, it is not an easy task to identify accurate positions for the WGWs. We have embedded into COP a highly accurate simulation model based on digital elevation and digital surface model with a resolution of 1/10 of an arc second. This allows visualizing together with the tactical information also the communication feasibilities. Both used together allow an accurate positioning of the WGWs within the operational area, which allows the field commanders to fulfil their tactical needs and being able to communicate to each other. More details about the simulation can be found in [16].

Figure 2 shows such an example output. The simulation output shows clearly where it will be possible to install a WGW and where it is not possible, based on the simulation result for a remote position.

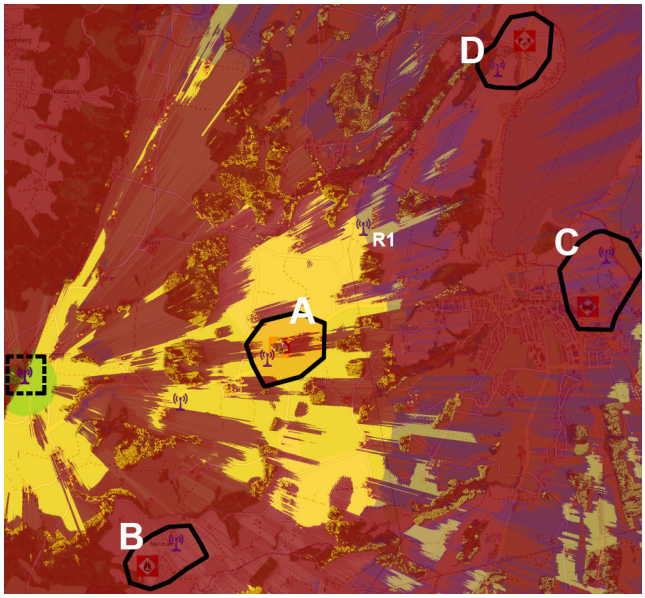


Figure 2. Simulation results.

The figure shows four incident areas (A-D) where communication coverage is needed. A central WGW has already been installed on the left border of the picture showing its simulation results. The lurid green color indicates the area where communication between end devices and the WGW will be possible. The yellow color shows areas where the installation of a remote WGW would be possible. The simulation results shows that for incident area A, the positioning of a WGW would be possible, but not for B, C and D. Consequently in a next step, a WGW will be positioned within area A and a simulation will be performed to show the coverage of the WGW positioned in that area. The aim is to find locations at the areas B, C, or D which are reachable from the WGW at location A. In doing so, it is possible to connect incident areas over multiple hops to the central WGW. For example, at position R1 on the map, a WGW is used as relay node to build a multi-hop-connection to incident area C.

The self-alignment algorithm of the WGW together with the simulation model allows the positioning and installation of the WGW by non-IT experts, thus the solution fulfills REQ4.

To be flexible to also integrate other technologies each WGW is using a so called Communication Field Relay (COFR). The Communication Field Relay is positioned at the foot of the pole on which the WGW is mounted, and it is connected to the WGW by Ethernet LAN. Furthermore, the COFR offers the ability to connect wired end devices such as desktop computers to the network. It provides all the needed networking services such as DHCP or DNS server for the machines. This link can also be used to integrate other layer 2 technologies, such as WIMAX, to the network. An additional link of the COFR is configured as DHCP client. On this link, different Internet uplink technologies such as DSL, Cable, 3G, 4G or anything else using an Ethernet interface can be connected. This Internet uplink can be shared by all clients connected to this COFR, to the local WGW or, to any remote COFR or WGW. The route is distributed by the OLSR dynamic gateway plugin.

Beyond networking, the COFR has a responsibility as power supply for the COFR/WGW compound. The COFR can be connected either to a 230V power socket or, to be independent of an available and working power grid, to a battery via a 12V cigarette lighter socket. The COFR provides power to the WGW via the PoE enabled Ethernet wire.

The IDIRA system consists of two more components the so called Fixed Infrastructure and the Mobile Integrated Command and Control Structure (MICS). From a networking perspective, the Fixed Infrastructure operates an OpenVPN server. As the COFR and WGW should work in different network environments two VPN configurations are used. One is using the default setting of an UDP VPN server on port 1194. The second one (overcoming some firewalls) is using a TCP VPN server running on Port 443. Moreover, the Fixed Infrastructure is running a DNS server. All the COFRs using an Internet uplink to connect to the IDIRA network will establish a VPN connection to the Fixed Infrastructure.

 ISBN: 978-1-61208-398-8

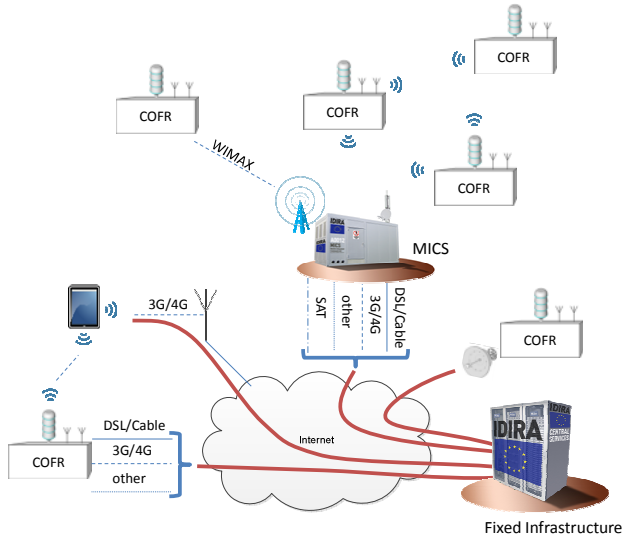


Figure 3. IDIRA communication network

The MICS allows shipping all the IDIRA services on site. This is of importance if a major network disruption occurs. When using the MICS, no Internet uplink is needed as all the services are running inside the MICS. All WGWs and COFRs will connect to the MICS. In the case that an Internet uplink can be established the MICS will setup a VPN connection to the Fixed Infrastructure and consequently, also hosts using the Internet can access the MICS. The flexible approach which allows using also existing networks to connect the COFRs to the Fixed Infrastructure and using other layer 2 technologies to interlink two COFRs ensure that also still existing network parts can be easily integrated into the network (REQ2). Figure 3 shows a deployment of the IDIRA communication network.

The proposed solution offers a speed of several Mbit/s which is sufficient for the needs of IDIRA (REQ7).

To fulfill REQ5 and REQ6 different steps on different components have been designed. When the system is fully operational, all the services can be accessed also including external expert systems which are running somewhere in the Internet. If the clients are only able to access the MICS, all services will be accessible, only the access to external expert systems will not be possible. When the connection between a COFR and the MICS is lost, the COFR will provide static information such as a map. Furthermore, all the information that has been viewed by a user before the network failure has been cached at the COFR and can be accessed. Also Voice over IP calls between end devices connected to the same COFR will still be possible. Finally, a native Android App has been developed to be used by the operational staff. All the data is synchronized between the MICS/Fixed Infrastructure and the Android App, so when the network is down all the information is still accessible by the users.

V. RESULTS

The presented communication solution has been used during several training actions and large scale exercises within the IDIRA project. Here, we will describe the setup used during two large scale exercises. The first exercise represents a disaster response operation after intense rainfall conditions. As a result, it came to severe flooding around the city of Görlitz at the Polish/German border. Due to the heavy rainfall, it came to landslides and building collapses. The exercise consisted of two parts - incident 1 and incident 2. At the incident 1 location, several people had to be rescued from the water. At the incident 2 location, (several kilometers apart from incident 1) some people had to be freed, as they were buried under a collapsed building.

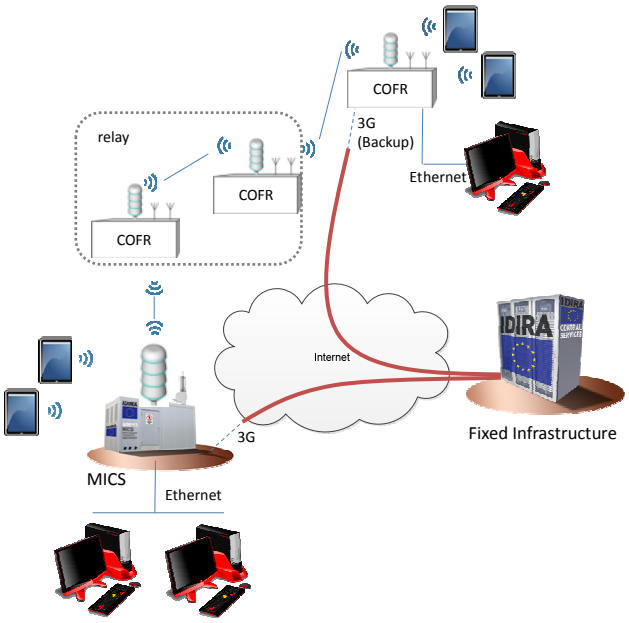


Figure 4. Real world deployment for first large scale exercise

Figure 4 shows the deployed infrastructure for the first large scale exercise. Incident 1 took place in close proximity (some 100m) to the on-site deployed MICS infrastructure. To provide connectivity to mobile end devices used at incident 1, a Wireless Gateway was used as access point and wire-connected directly to the MICS. Several workstations were connected to the MICS via Ethernet LAN.

Also at the location of incident 2, mobile end devices should also be connected to the MICS infrastructure via wireless LAN and a WGW access point. The problem here was, that incident 2 took place at a location 3.5 km distant from the MICS installation, separated by hills, dense forests and even urban area. To span a point-to-point WLAN connection over more than 3 kilometers, and to bypass obstacles (hills, trees, buildings, etc.), it was decided to install two relay nodes to provide the needed communication coverage at the incident 2 location. All four WGW nodes – the one at the MICS location, the two relay nodes, and the one at the incident 2 location - were separated by distances

 ISBN: 978-1-61208-398-8

between 200m and 1.3 km. Except for the WGW located at the MICS, all relay nodes were installed by a unit of the Austrian Red Cross.

In addition, at location of incident 2, a mobile command and control vehicle was connected to the MICS using the Ethernet LAN connection offered by the COFR. The connection was used to access and operate the COP.

At location of incident 2, we decided to prepare an additional 3G uplink. This uplink should provide backup connectivity in case of a lost WGW point-to-point connection and could have been used to connect mobile devices over 3G to the MICS Infrastructure - even though with a very limited bandwidth and performance. These bandwidth and performance limitations are further exacerbated, as the MICS itself uses only a 3G Internet uplink due to a lack of availability of other broadband connections like DSL.

Both, the 3G uplink and the WGW point-to-point connection were tested during the exercise.

Figure 5 shows the deployment as it has been used during the second described large scale exercise. The exercise scenario was about an earthquake and fire disaster. As a result of the earthquake (which had its' epicenter in the area of Attica, Greece), it came to multiple blocked roads, collapsed buildings and fires. The exercise consisted of four incident areas, which were spread over an area of several kilometers in square.

In contrast to the first exercise, where end devices have been connected only via WGWs and 802.11, at one incident location, tablet devices have been equipped with 3G SIM cards to use the existing 3G network. Similar to the backup solution at incident 2 in the first exercise, we also used a 3G uplink for the COFR/WGW compound at one incident location. In contrast to the first exercise, the on-site deployed MICS could be connected to a DSL uplink. Several PCs have been directly connected to the MICS via Ethernet. A remote PC was using a legacy WIFI network which is normally used for cameras, and a COFR/WGW network has been established to be fully independent in two disaster areas. Also the Offline functionality with the Android App has been tested by switching off the data connection on the tablet device using the SIM card.

Again, the installation of the wireless communication infrastructure was done by a unit of the Austrian Red Cross.

The deployments in both exercises have shown the flexibility of our solution in combining existing networks with COFR/WGW networks (REQ2).

All over the exercises the communication infrastructure was installed by different personnel of various first responder organizations. The setup of the system has been proven to be easy and can be fulfilled by non-IT experts (REQ4). During the trainings it was shown, that battery, pole, tripod, COFR and WGW were able to be transported and installed by one single person.

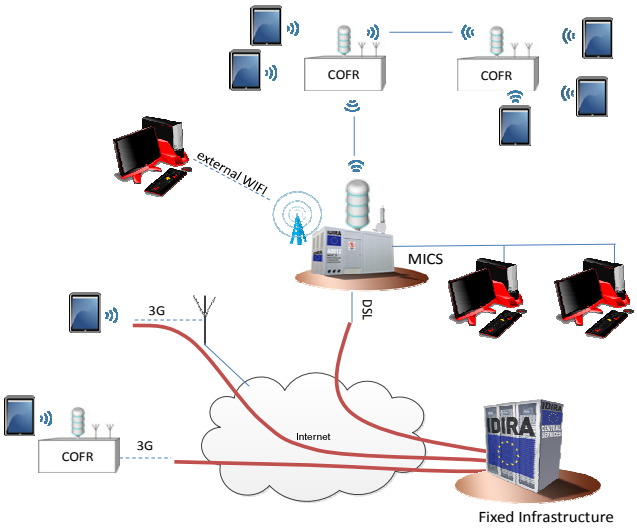


Figure 5. Real world deployment of the IDIRA communication network

Furthermore, the assessment of the communication system was part of surveys, which collected the views of all relief forces using the IDIRA system. A majority of those questioned about the communication infrastructure are of the opinion that the communication solution would be helpful in real deployments and the communication quality and coverage was (very) sufficient for the use with the IDIRA system.

VI. Conclusion And Future Work

This paper presents a flexible, communication solution for large scale multi-national and multi-organizational relief operations. The concept complies with several requirements which have been introduced by action forces of relief organizations, such as easy installation and transportation, interoperability with existing communication systems and, international operation permission. The core of the system consists of the WGW/COFR compound. The WGW are able to automatically build up meshed wireless networks, using self-aligning directional antennas. The COFR provides power supply and Internet uplink technologies to the compound. During several trainings and large scale exercises, the system was able to prove its workability for the use within the IDIRA system.

For the future it is planned, to mechanically enhance the prototypes to fulfil the mechanical requirements of robustness for being used in real world large scale disasters.

Furthermore, we plan to integrate a MANET based approach similar to the one described in [15] to expand the communication coverage created by the proposed WGW solution.

 ISBN: 978-1-61208-398-8

ACKNOWLEDGMENT

This work was partially supported by the IDIRA European FP7 261726 research project.

REFERENCES

[1] IDIRA Project. *Interoperability of data and procedures in large-scale multinational disaster response actions, 2011-2015* [Online]. Available from: http://idira.eu/. 2015.02.19

[2] Immarsat BGAN. *Broadband Global Area Network* [Online]. Available from: http://www.inmarsat.com/service/bgan/ 2015.02.19

[3] GVF VSAT. *Global Very Small Aperture Terminal Forum* [Online]. Available from: http://www.gvf.org 2015.02.19

[4] Emergency.lu. [Online]. Available from: http://www.emergency.lu 2015.02.19

[5] IEEE 802.16 WIMAX. *IEEE Standard for Local and metropolitan area networks* [Online]. Available from: http://standards.ieee.org/about/get/802/802.16.html 2015.02.19.

[6] IABG mbH. *HiMoNN Higly Mobile Network Node* [Online]. Available from: http://www.himonn.de 2015.02.19

[7] Electronic Communications Committee (ECC). *The Identification of Frequency Bands for the Implementation of Broad Band Disaster Relief (BBDR) Radio Applications in the 5 GHz Frequency Range* [Online]. Available from: http://www.erodocdb.dk/docs/doc98/official/pdf/REC0804.pdf 2015.02.19

[8] IEEE 802.11. *IEEE Standard for Information Technology--Telecommunications and Information Exchange Between Systems--Local and Metropolitan Area Networks--Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY)* [Online]. Available from: http://standards.ieee.org/about/get/802/802.11.html 2015.02.19

[9] T. Clausen and P. Jacquet. (2003), "Optimized Link State Routing Protocol (OLSR)", Internet Engineering Task Force, IETF, RFC 3626.

[10] OASIS Emergency Management TC. *Emergency Data Exchange Language (EDXL)* [Online]. Available from: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=emergency 2015.02.19

[11] OASIS Emergency Management TC. *Common Alerting Protocol Version 1.2, 2010-07-01 - CAP-v1.2-os* [Online].Available from: https://www.oasis-open.org/standards#capv1.2 2015.02.19

[12] OASIS Emergency Management TC. *Emergency Data Exchange Language Resource Messaging, EDXL-RM-v1.0-OS-errata-os, 22 Dec. 2009* [Online]. Available from: https://www.oasis-open.org/standards#edxlrm-v1.0 2015.02.19

[13] OASIS Emergency Management TC. *Emergency Data Exchange Language Situation Reporting, edxl-sitrep-v1.0-wd19, Draft 02, 2012-08-07* [Online].Available from: http://docs.oasis-open.org/emergency/edxl-sitrep/v1.0/cs01/edxl-sitrep-v1.0-cs01.zip 2015.02.19

[14] P. Dorfinger, G. Panholzer, F. von Tüllenburg, M. Christaldi, G. Tusa, and F. Böhm, "Self-aligning Wireless Communication for First Responder Organizations in Interoperable Emergency Scenarios", Proc. of the 2014 International Conference on Wireless Networks (ICWN 2014), July 2014, pp. 151 – 157, ISBN: 1-60132-278-X.

[15] C.Raffelsberger and H. Hellwagner, "Evaluation of MANET Routing Protocols in a Realistic Emergency Response Scenario" Proc. of the 10th Workshop of Intelligent Solutions in Embedded Systems (WISES'12). July 2012, pp. 88-92.

[16] T. Pfeiffenberger, P. Dorfinger and F. von Tüllenburg, „Communication Coverage Awareness for Self-Aligning Wireless Communication in Disaster Operations" Proc. of the 5th International Workshop on Perversave Networks for Emergency Management (PerNEM2015). In press.

 ISBN: 978-1-61208-398-8

Interference-Aware Routing Supporting CARMNET System Operation in Large-Scale Wireless Networks

Maciej Urbański, Przemyslaw Walkowiak, Pawel Misiorek
Institute of Control and Information Engineering
Poznan University of Technology
M. Sklodowska-Curie Sq. 5, 60-965 Poznan, Poland
Email: {maciej.urbanski, przemyslaw.walkowiak, pawel.misiorek}@put.poznan.pl

***Abstract*—The paper provides the research results concerning interference-aware routing metrics supporting the operation of large-scale wireless mesh/multi-hop networks. In particular, the description of state of the art in this domain of research has been provided and used in order to define a new interference-aware metric for multi-radio networks. The proposed solution is aimed at supporting traffic control mechanisms for multi-hop wireless networks by improving performance of any system controlling multi-hop wireless transmission, e.g., a system widening the access to the Internet by means of wireless communication such as a CARMNET system. Moreover, the paper presents the implementation of a set of tools necessary to provide experimental evaluation of the solution in a large-scale wireless environment as well as a set of tests comparing the performance of the proposed metric with state-of-the-art solutions. The experiments have been performed in the DES-Testbed located at Freie Universität Berlin, which is one of the largest wireless testbeds in Europe. Each of the nodes taking part in experimentation has a CARMNET Loadable Linux Kernel Module (CARMNET LLKM) deployed.**

Keywords–wireless mesh/multi-hop networks; interference-aware routing; large-scale wireless experimentation.

I. INTRODUCTION

Wireless networks are more and more popular due to the rise of mobile devices. While their capacity is much lower than capacity of their wired counterparts, they are highly valued by end users for ease of setup and unrestrained mobility. Mesh networks technologies, however, while very promising in theory are still underused by industry due to their complexity. In order to operate correctly, robust mesh networks require additional protocols for dynamic routing, resource allocation, and self-configuration.

The presented research has been motivated by the need for routing optimization necessary to improve the performance of a CARMNET resource management system operation [1] in large wireless networks. With the development of Wireless Mesh Network (WMN) the importance of routing protocol becomes more and more apparent. The interference-aware routing metrics which take advantage of wireless networks' properties, are especially important for the WMN researchers. One of the weaknesses of a multi-hop mesh network is self-interference, because of which, even in a scenario with a single transmitting node, the available bandwidth is much lower than that achievable throughput on the single link. More complex WMNs mitigate this problem by incorporating multi-radio nodes, allowing multiple transmission to occur at the same time without interference. For such networks, the development of special routing path metrics which take advantage of the non-interfering channels has to be provided. It has to be stressed that for such solutions interference range is still vague and certainly not limited to the 1-hop neighborhood. Moreover, the metrics have to provide the trade off between their complexity and achievable gains to make them practical enough for the implementation in real-world scenarios.

The goal of this paper is to provide the routing solution for large-scale multi-hop multi-radio wireless networks which is suitable to support the operation of resource management mechanisms based on utility maximization. In particular, the paper presents the experimental research results concerning the test on routing metrics performance conducted in large wireless testbed with nodes controlled by utility-aware resource management subsystem referred to as CARMNET Loadable Linux Kernel Module (CARMNET LLKM) [1] based on DANUM subsystem [2]. The module is a part of CARMNET system developed by the research team realizing the Polish-Swiss Research Programme project CARMNET "CARrier-grade delay-aware resource Management for wireless multi-hop/mesh NETworks" [3] devoted to research on delay-aware wireless networking, multi-criteria routing, and the IMS (IP Multimedia Subsystem) reliable application in a wireless environment. Although the provided research on interference-aware routing is motivated by the need of optimizing the CARMNET system operation in mesh/multi-hop networks, it may be also beneficial for the research on other aspects of multi-radio multi-hop wireless networking.

The rest of the paper is structured as follows. Section II provides a discussion on related work, which contains a brief presentation of state-of-the-art interference-aware routing metrics. The introduction of the CARMNET framework is given in Section III and the proposal of a new interference-aware routing metric is presented in Section IV. Then, Section V is focused on the experimental research. It contains the description of the testbed environment, the implementation of tools necessary to conduct the tests, experimentation scenarios, results and their analysis. Finally, the paper is concluded in Section VI.

II. RELATED WORK

The presented research on routing metrics has been motivated by the need to optimize the wireless resource allocation solutions aimed at maximizing network utility. Many resource allocation systems based on the Network Utility Maximisation (NUM) model exist and determine the utility of flows according to their measured properties [1][2]. However, only a few of them have been implemented and tested in a real-

 ISBN: 978-1-61208-398-8

world wireless mesh network [4][5]. Furthermore, the proposed approaches are not sufficient to effectively measure the utility of both delay-sensitive and throughput-oriented flows. On the other hand, the CARMNET system [1] uses a CARMNET Loadable Linux Kernel Module as a resource management subsystem based on DANUM System (DANUMS) [2], which takes both parameters into consideration. Moreover, the CARMNET system has a well-tested implementation [1], which allows researchers to focus on specific parts of the system operation.

The remaining part of this section is aimed at introducing the state-of-the-art interference-aware routing metrics used for the comparison presented in this paper. In general, the presence of interference is one of the most characteristic features of the wireless networks and a major factor constraining their performance [6]. Depending on a source of the interference, it can be categorized as controlled internal or external interference. Controlled internal interference can be reduced by the modification of the network properties (channel assignment, routing, scheduling) and can be further divided into inter-flow and intra-flow interference. Inter-flow interference may be described as a harmful competition for medium between routers when transmitting multiple flows. Intra-flow interference occurs when transmitting the single flow over a multi-hop wireless path, for which flow transmission rate is radically reduced since the medium has to be shared between each hop of a transmission. In the real-world scenario, external, uncontrolled interference in wireless systems has to be expected and should be taken into account during the routing performance analysis. Depending on a source, it can be more or less predictable and dynamic. The 802.11 networks use the Industrial, Scientific and Medical (ISM) radio bands, which are also applied in other technologies, such as Bluetooth, ZigBee or proprietary wireless audio systems.

A. Expected Transmission Count

With hop count metric proven to be ineffective in irregular WMNs topologies [7], the Expected Transmission Count (ETX) became the most popular metric for the WMN. The value of the ETX metric for a bidirectional link is calculated as it shown in the following formula:

$$ETX = \frac{1}{(1 - P_f)(1 - P_r)}, \quad (1)$$

where P_f and P_r are probabilities of the packet loss when transmitting in the forward (i.e., $node_A \rightarrow node_B$) and reverse (i.e., $node_A \leftarrow node_B$) direction, respectively.

During the implementation of the ETX metric (and each metric based on it), it is crucial to consider how a packet loss is handled by lower and upper layers (e.g., it is crucial to take into account if there is a retransmission mechanism applied).

B. Expected Transmission Time

Expected Transmission Time (ETT) is an extension of the ETX metric based on introducing a link speed factor. For each link, a value of the metric is computed according to the following formula:

$$ETT_l = ETX_l \frac{S}{B_l}, \quad (2)$$

where S represents size of the packet and B_l represents the link l data rate (as indicated by Link layer).

The ETT value can be seen as inversely proportional to so called 'goodput' of the link, representing successful packet delivery rate. The most of the implementation use only probes of the arbitrary size, which is not tied to S parameter. If all link costs are multiplied by the same constant value of S, the S parameter becomes entirely insignificant for the task of the path ordering.

C. Weighted Cumulative ETT path metric

Weighted Cumulative ETT (WCETT) [8] is built directly on the ETT metric, with the aim to achieve better performance over multi-radio links. The following formula describes the WCETT path metric for path p using K non-interfering channels [8]:

$$WCETT_p = (1 - \beta) \sum_{l \in p} ETT_l + \beta \max_{1 \leq j \leq K} X_j, \quad (3)$$

where $X_j = \sum_{l \in p^j} ETT_l$ is the attainable throughput in the single channel. The subset p^j of links on path p is defined as $p^j := \{l : l \in p \wedge channel(l) = j\}$, where function $channel(l)$ returns a channel which link l is associated with.

As it can be concluded from (3), the X_j value of the channel, which represents the bottleneck of the path (with the maximum value), is taken into account in the measure. The β value, such that $\beta \in \langle 0, 1)$ controls channel diversity. For $\beta = 0$, WCETT becomes identical to the ETT metric. Setting $\beta = 1$ is not recommended as in such a case, the metric treats longer path (in terms of the hop count) as identical as far as they use non-interfering channels.

D. Additional metrics

The performance evaluation results presented in this paper are reduced to above-mentioned metrics. However, this set may be extended by several other metrics. In particular, the Metric of Interference and Channel-switching (MIC) [9] is a metric which is based on heuristic weighting of two parts addressing the inter-flow interference impact and intra-flow interference impact, respectively. The Exclusive Expected Transmission Time (EETT) metric [10] is another interference-aware routing metric which tries to solve the problem of performance degradation in the large-scale WMNs. Finally, Interference Aware (iAWARE) metric [11] is the metric which, in the contrary to above mentioned interference-aware metrics, takes into account the physical interference model and uses signal strength of heard or sensed packets to determine Signal to Interference and Noise Ratio (SINR).

Table I summarizes the comparison of state-of-the-art routing metrics by presenting, which metrics consider particular aspects of routing in the WMN including *link loss*, *link speed*, *intra-flow interference* and *inter-flow interference*, *isotonicity*, and awareness to the *multi-radio* transmission.

Depending how the link metrics calculate the overall path cost, they could be further divided into the (i) additive metrics – the path cost is the sum of the metric value of all links, (ii) multiplicative metrics – the path cost is the product of the metric value of all links, and (iii) statistical metrics – the path cost it the minimum/maximum/average of the cost

 ISBN: 978-1-61208-398-8

TABLE I. FEATURES OF POPULAR WMN METRICS.

	link loss	link speed	intra-flow interf.	inter-flow interf.	isotonicity	multi-radio
hop count	no	no	no	no	isotonic	no
ETX	yes	no	no	no	isotonic	no
ETT	yes	yes	no	no	isotonic	no
WCETT	yes	yes	yes	no	monotonic	yes
MIC	yes	yes	yes	yes	no	yes
EETT	yes	yes	yes	yes	isotonic	yes
iAWARE	yes	yes	yes	yes	no	yes

of the individual links. The metrics discussed in this paper are mostly additive (hop count, ETX, ETT), what basically is an assumption used by the majority of the routing protocols.

III. CARMNET FRAMEWORK

The assumptions and architecture of the CARMNET system have been proposed with details in [1]. The main goal of the system is to enable the WMN users to share their resources, in particular to share the Internet access. Additionally, the CARMNET vision assumes the work on integration of the delay-aware resource allocation subsystem with the Internet provider infrastructure [12]. CARMNET solutions are aimed to be integrated with public wireless networks — their have been already tested in municipal network WiFi Lugano [13] or in the socially-operated network – Malta NET – located in Poznan [14]. The Internet sharing functions (described with details in [12]) are optimized as a result of the application of the utility-aware resource allocation subsystem (i.e., the CARMNET LLKM based on Delay-Aware Network Utility Maximisation (DANUM) model [2]), which allows to compare the utility of flows with different requirements with regard to end-to-end delay and throughput.

The implementation of the CARMNET framework requires integrated studies in several research areas including wireless network resource management, multi-criteria routing, and seamless handover [15]. In this paper, we have focused on issues concerning the routing solutions supporting CARMNET resource management subsystem, i.e., DANUMS, which is responsible for resource allocation.

A. CARMNET LLKM

The CARMNET LLKM resource management subsystem (which is based on DANUM system [2]) is aimed to maximise the overall utility of a mesh network defined as:

$$\sum_{r \in S} U_r(x_r, d_r), \tag{4}$$

where S denotes a set of flows within the network; x_r – rate of flow r; d_r – delay of flow r; U_r – the utility function of flow r [2].

In order to optimise the allocation of network resources, DANUMS prioritizes flows, which gain the most utility from being served. The solution is based on "virtual queue" levels defined as a product of a packet backlog and the value of the first derivative of a given flow utility function calculated for current flow performance parameters [2]. The parameters which influence the utility value include flow's packets delivery delay and throughput measured at destination node. Both the mentioned parameters are attained by the active measurement through the use of Delay Reporting Message (DRM) [2]. In parallel, virtual queue levels are used to perform Max-Weight Scheduling (MWS) [16] on each hop in distributed manner. The details of DANUMS implementation may be found in [1] [2].

IV. DESIGNING THE WCETTX PATH METRIC

The goal of the proposed WCETT-eXtended (WCETTX) metric is to improve the WCETT metric, which is based on ETT optimized in a way enabling simple interference-awareness for the multi-radio networks, and in consequence, to define the metric suitable to be used in CARMNET-controlled networks.

The WCETT metric assumes that by using several channels one can avoid a part of interference. Following this assumption, WCETT focuses on the maximum additive sum of ETT links' values, grouped by the channel used. This sum represents the 'bottleneck of the path' and is recognized as a major contributor to its cost. Still, the authors of WCETT acknowledge that additional hops on the channels, which are not considered as a bottleneck, represent the additional cost. It is modeled by an additional sum of overall links' ETT values, which is balanced with the major part of the metric by arbitrarily assigned constant. The setting of this constant is problematic, since one configuration cannot accommodate all of the scenarios [8]. The aim of WCETTX is to address this problem and to propose a more natural representation of the additional hops cost.

A. Formulation of WCETTX

The main improvement of WCETTX over WCETT is a clear representation of the additional hop cost of links on non-interfering channels. The assumptions made when formulating the WCETTX metric are as follows:

1) The metric of link throughput on interfering links is used in the additive manner, as transmissions on links 'divide' available transmission time;
2) The metric of network layer loss occurring on each hop is used as multiplicative one;
3) The channel with the maximum additive throughput metric value represents a path bottleneck and is regarded as the main factor limiting possible throughput;
4) The cost of the additional hops can be represented as the loss risk;
5) In the real-world scenario, the perfect links do not exist and a loss rate is always higher than zero, (i.e., $\underset{\text{link } l}{\forall} ETX_l > 1$).

The throughput link metric, such as ETT, representing 'goodput' of link, is commonly used as additive. Multiplicative use of the loss-based metric is not so often seen in wireless networks, as they implement a retransmission mechanism on the Link layer. However, even this mechanism fails sometimes, and loss can be seen on the higher layer in end-to-end matter, what justifies its use as the multiplicative metric. Assumption 3 is shared with the WCETT metric – each used channel is treated as not having any impact on other channels. Assumption 4 represents an optimization over WCETT aimed at differentiating between shorter and longer paths without the need for additional arbitrary assigned variables. Finally,

 ISBN: 978-1-61208-398-8

assumption 5 is used to prevent longer (in terms of hop count) paths to be treated as equal due to the close to ideal environment and measurement inaccuracies.

Equation (5) defines the WCETTX path metric:

$$WCETTX_p = \max_{1 \le j \le K} X_j \times \prod_{l \in \mathrm{p}} \frac{1}{1 - netw\_layer\_loss_l}, \quad (5)$$

where K is a number of channels, $l \in$ p denotes links of path p, and X_j is defined as in the case of the WCETT metric. The $netw\_layer\_loss_l$ equals to $\left(1 - \frac{1}{ETX_l}\right)^{retry}$ and is used to describe the cost of additional hops.

Probability of loss is different from Network- and Link-layer perspective, as the lower layer can retransmit the packet preventing the higher layer loss. Such a retransmission mechanism is implemented in Wi-Fi networks if no acknowledgment for the transmitted packet is received. If loss occurs, the packet is regenerated, as long as it is under $retry$ limit of retransmission attempts. This mechanism makes Network layer loss much less likely, but still possible. This end-to-end loss is applied to estimate path throughput, here represented as X_j, in the similar way like it is done in the case of the ETT metric.

B. The analysis of the WCETTX metric

The objective of this subsection is to discuss the requirements for the routing protocol necessary to apply the proposed WCETTX metric.

First, the WCETTX metric assumes that measurements of both link's loss and throughput have to be used instead of measurement of a single parameter. This approach may be regarded as an improvement over the WCETT metric, since it allows to separately model the influence of these parameters on the path cost. Additionally, it enables to take into account the association of the link to the interfering or non-interfering channel. The third parameter that must be monitored for each advertised link is a channel on which it operates. This information can be inferred by the neighboring nodes, but has to be explicitly announced during further forwarding of the topology information.

Similarly as WCETT, the WCETTX metric is non-isotonic, but monotonic. This feature has its consequences in both the routing path calculation process and routing itself. In the most common hop-by-hop routing scheme used in networks based on TCP/IP, the path is determined according to the destination node only. For non-isotonic metrics, such as WCETTX, the paths are also source dependent, thus, a more complex process for their computation is required. Moreover, the well-known algorithms used for routing path calculation, such as Dijkstra's algorithm or Bellman-Ford algorithm, process nodes in breadth-first manner and can only be applied with isotonic metrics [17]. The authors of [8] have noted that the simple adaptation of the k-shortest paths algorithm provides non-optimal routes in terms of the non-isotonic WCETT metric. The reason for that is the fact that the maximum function is used for metric calculation, what is the case also for the proposed WCETTX metric.

It has to be admitted that due to the requirement of data gathering over time as well as computational difficulty of path calculation (which slows down the time of reaction to change in the network), the WCETTX metric seems to be suited for static WMNs only. In order to use the metric in the Mobile Ad hoc Network (MANET), the ability of fast path recalculation based on a previous path should be further investigated.

It should be stressed that WCETT concentrates on intra-flow interference, while not counteracting inter-flow interference. This feature makes this metric more suitable for networks in which only several nodes exchange information between each other. Following this assumption it is worth considering to calculate WCETTX paths on-demand, only for nodes exchanging high loads of traffic, while using underlying ETT metric for fast path calculation proactively for all the nodes.

V. Experiments

The physical testbed environment has been chosen for the purposes of the proposed solution testing. While simulations offer better control over all experiment parameters as well as better scalability, these advantages are always achieved at some cost – in the simulation environment the link and physical layers of the Open System Interconnection Reference Model (OSI-RM) are often simplified, what in the case of testing interference-aware solutions could severely affect the results.

A. Testbed

The Distributed Embedded Systems Testbed (DES-Testbed) [18] has been used in the experimentation described in this paper. DES-Testbed is a non-commercial testbed designed for the purposes of research in the area of wireless mesh and sensor networks. The testbed is divided into two networks, the first one being WMN, called DES-Mesh, and the second one called DES-WSN being sensor based. Both parts are connected, as DES-Mesh consists of the core nodes, and DES-WSN is based on daughter-boards integrated with the core nodes. The whole DES-Testbed consists of over 100 core nodes, each integrated with a daughter sensor node, what makes it one of the largest academic wireless testbed. The core nodes are based on the x86 embedded PC boards, each equipped with up to three wireless network adapters. At the time of conducting the tests described in this paper only half of nodes were available for testing, since the part of DES-Testbed and the team responsible for it have moved to the University of Münster.

The DES-Testbed nodes are placed around Freie Universität Berlin (FUB) campus, thus the network topology represents a real-world scenario in which placement of nodes is irregular and external interference sources exists (e.g., generated by actively-used other networks). By default, each of the DES-Testbed nodes connects (by means of its three interfaces) to three network cells: *des-mesh0*, *des-mesh1*, and *des-mesh2*. Each of these network cells is configured to work at a separate channel, in order not to interfere with other cells. However, during experimentation we have noticed that even when using separate channels, the interference between radios may be observed. This problem is similar to the one encountered by authors of [8]. In order to mitigate the multi-radio interference issue, only one channel from of 2.4GHz band and one channel of 5Ghz band were used during experimentation. This setting corresponds to interface "wlan0" and "wlan2" of DES-Testbed nodes (des-mesh0 and des-mesh2, respectively).

Figure 1 shows the complexity of network cells *des-mesh0* and *des-mesh2*.

 ISBN: 978-1-61208-398-8

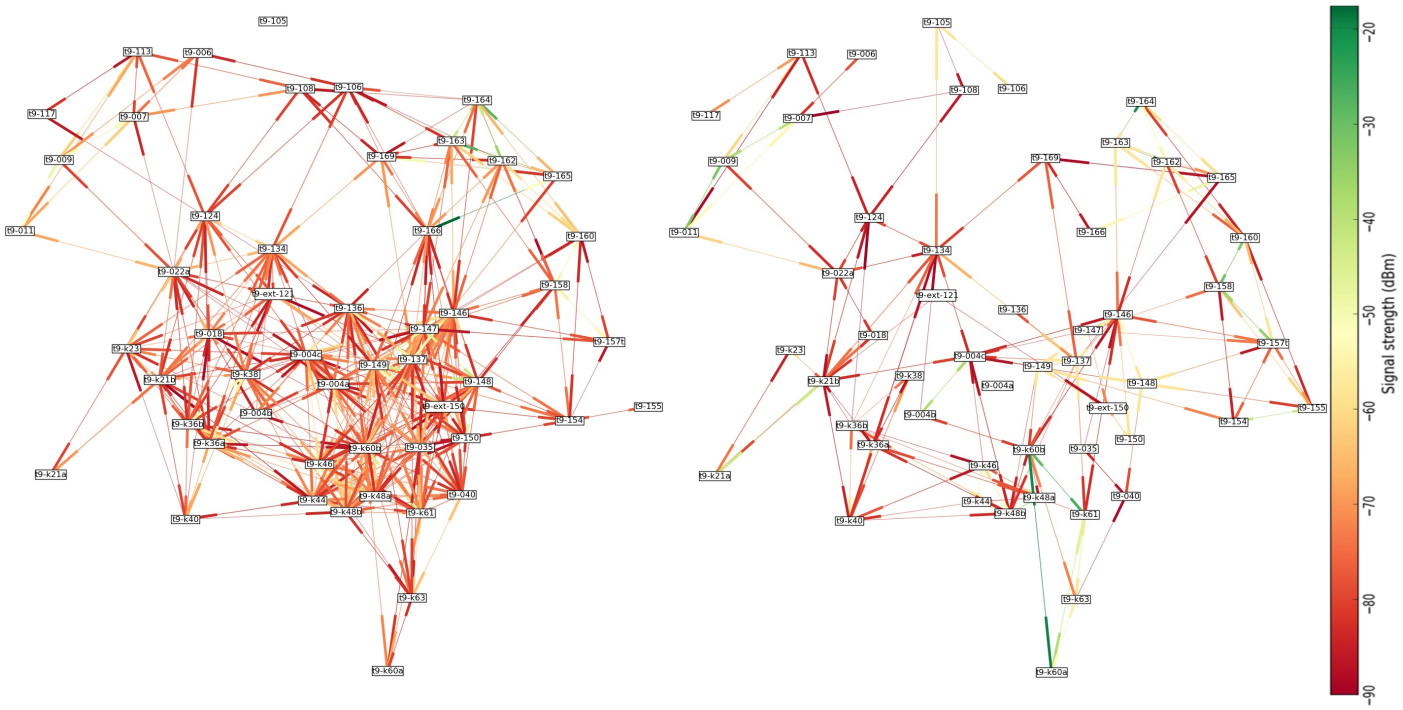


Figure 1. Overview of the signal strength in DES-Testbed network cells (from the left: *des-mesh0*, *des-mesh2*).

B. Implementation

This section contains the description of tools implemented by the authors, as well as provides the motivation for the choice of specific solutions related to technical aspects of networking in Linux, which is integral part of the testbed used.

1) Source-determined routing in Linux operating system: Although source-determined routing is required for a number of the interference-aware routing metrics, its implementation in Linux is not available. There are two ways to force the packets to travel between source and destination along a specific path in a TCP/IP network. The first one (later referred to as a Source routing IP option) is to use IP packet options Strict Source and Record Route (SSRR) or Loose Source and Record Route (LSRR) as specified by [19]. The second one (referred to as policy routing) is to make each node to choose a particular path based on both source and destination IP addresses and the set of predefined rules. Source routing IP option and source routing in the form of policy routing are easily mistaken and hard to investigate due to similarity of terminology and unpopularity of the former. The Linux kernel networking stack supports both approaches, but to a different extent. Linux user space policy routing utilities allow configuration of the routing policy rule based on a source address of the packet, which is later executed during the routing process. While in the case of the Source routing IP option, the Linux Kernel implements its handling, respecting both LSRR and SSRR headers, there are no utilities allowing insertion of such headers into the packet. Such headers could potentially be implemented with the use of the Netfilter framework or the network tunnel (TUN) virtual device, but such a solution would require introduction of additional protocol overhead and more computation as IP packet had to be created anew. The shortcomings of the source policy routing also exist, since for this approach, each possible source node has to be filtered to the separate routing table. The hard limit of the routing tables number in Linux is less than 256, which can be a problem for very large networks.

After considering all of the pros and cons, the source policy routing has been chosen in our implementation. Main factor for the decision was maturity and stability of this solution, which is important for the remote experimentation on a physical testbed, for which a recover from the potential kernel-level crash is hard to be done without a direct intervention.

2) Topology measurements and path calculation software: For the topology mapping, the basic Linux networking tools like *iw* and *iperf* have been used. The *iw* – the wireless networking configuration tool – was used to list each node neighbors detected at the Link Layer (in this case 802.11). The User Datagram Protocol (UDP) transmissions were generated by *iperf* on the source node with the maximum possible rate (as chosen by the Link layer rate adaptation mechanism) and were measured on the destination node, thus providing both packet throughput and loss rate for the interconnecting link. To ensure the reliable measurement of Link layer packet loss rate, the retry mechanism was disabled using *iwconfig* utility.

For the additional processing of the topology graph, the *Networkx* python module was used. The process of topology mapping and path calculation consisted of following steps (i) measurement gathering and generation of the network graph (ii) finding alternative paths between each pair of nodes, (iii) computing metrics for alternative paths and choosing the best path for each metric, and (iv) preparation of policy routing rules.

C. Experimentation assumptions

The experimentation follows the methodology used by authors of the WCETT metric described in [8]. The objective

of the experimentation scenario was to verify intra-flow interference reduction properties of WMN routing metrics designed for multi-radio wireless communication.

The following metrics were tested: WCETTX, WCETT multi-radio metrics and the ETX metric used as a baseline comparison. In addition, for WCETT, various β values have been examined. The experiment has been run on 52 DES-Testbed nodes. The routing paths were precalculated for all of 2652 ($N * (N - 1)$, where $N = 52$) source-destination pairs beforehand, as described in Subsection V-B2. From these pairs, 200 have been selected at random in order to be used as sender-receiver pairs during throughput measurement tests. The experimental evaluation presented here is limited to the tests on these randomly selected 200 pairs of source and destination nodes. The Transmission Control Protocol (TCP) connections generated by *iperf* have been used to check achievable end-to-end throughput. The Linux default "CUBIC" congestion algorithm was used, which has much more aggressive slow-start phase [20]. This feature allowed to lower the connection time down to 20 seconds when estimating the path bandwidth. To ensure that the network is clear between each test, an additional second was spent waiting after a single test to guarantee that no more packets are queued.

D. Results and analysis

Due to the hard-to-locate bug in the implementation of the 802.11 stack, in the network cell association routines, the des-mesh network cells tended to partition. In order to mitigate its effect, the presented analysis has been limited only to the source-destination pairs for which all the provided test executions have finished successfully without being influenced by the above-mentioned bug. From 200 possible samples, no less than 112 samples were gathered in each test according to this criterion.

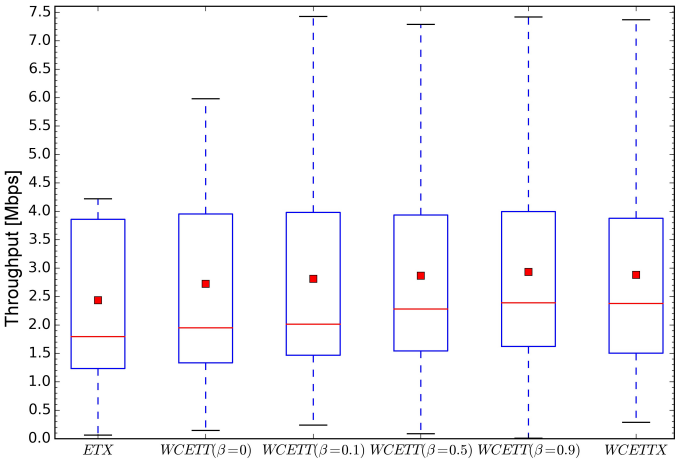


Figure 2. Two-channel scenario results obtained using *des-mesh0* and *des-mesh2* network cells.

Results of the experiment are based on 112 common samples of throughput measurements. Figure 2 is a box plot of achieved throughput. Table II repeats these results in a numerical fashion. The outliers are not shown in Figure 2, as they depict only single-hop bandwidth and are similar in every tested metrics.

TABLE II. ACHIEVED THROUGHPUT FOR EACH OF THE METRICS USING *des-mesh0* AND *des-mesh2* NETWORK CELLS.

metric	median	standard deviation	average
ETX	$1.79Mbps$	$2.26Mbps$	$2.44Mbps$
ETT	$1.95Mbps$	$2.68Mbps$	$2.73Mbps$
$WCETT(\beta = 0.1)$	$2.02Mbps$	$2.45Mbps$	$2.81Mbps$
$WCETT(\beta = 0.5)$	$2.28Mbps$	$2.37Mbps$	$2.87Mbps$
$WCETT(\beta = 0.9)$	$2.39Mbps$	$2.40Mbps$	$2.93Mbps$
$WCETTX$	$2.38Mbps$	$2.60Mbps$	$2.88Mbps$

The ETT (WCETT $\beta = 0$) metric has achieved significantly better results compared to ETX (the observed improvement of median throughput is over 18%), which is expected due to diversity of link quality. For $\beta = 0.1$ WCETT, the difference to the results of the standard ETT is not significant. For higher β values, WCETT delivers better throughput up to over 20% of improvement for median throughput when compared to basic ETT. WCETTX provides results comparable to WCETT having the β value optimized. In order to clarify the benefit from applying the proposed metric, it has to be stressed that, for a given network configuration and experiment scenario, the optimal value of β is not known in advance, and needs to be determined heuristically. The provided results confirm that in the case of WCETTX, the metric performance is equal to the maximum performance of the optimized WCETT and is obtained without the need of any parameter optimization.

It has been also noticed that even when using separate frequency bands the multi-radio interference persists. It is important to be aware that such a kind of interference may occur what may lead to the results for which the simple single-radio metrics outperform the interference-aware ones. The problem was mitigated when power of the radios was limited. The results presented here were obtained using $18dBm$ transmission strength setting.

VI. CONCLUSION AND FUTURE WORK

The main contribution of this paper is the experimental analysis of interference-aware routing metrics for multi-radio multi-hop wireless networks, which is devoted to support operation of resource management mechanisms in large-scale wireless networks. The evaluation has been performed in the DES-Testbed wireless environment using nodes operated by experimental utility-aware resource management subsystem based on CARMNET LLKM. The presented metric comparison includes the new metric proposed by the paper authors. The proposed WCETTX metric is based on the theoretical assumption of non-interference between channels of different bands and has been proven to provide the better or similar performance than other compared solutions without the heuristic optimization of parameters.

It has to be stressed that the tests in the real-world environment, while more time consuming, provide better picture of all the issues which could arise during the deployment of the solution. In particular, the tests have shown that the occurrence of multi-channel interference is possible. To the authors knowledge, no WMNs simulator implements the model which reflects performance degradation connected to the issue of the multi-radio interference.

The future work plan includes more detailed analysis of the multi-radio interference and its consideration during designing path metric. Additionally, we are going to conduct the

 ISBN: 978-1-61208-398-8

extended experimentation using various topologies of different characteristics aimed at additional comparison of WCETT and WCETTX metric routing performance. The goal of these new experiments is the analysis of limitations of both solutions and providing the additional evidence of WCETT inaccuracy in estimation of the additional hop cost, which has been observed in results provided in [8].

ACKNOWLEDGEMENT

This work was partly supported by a grant CARMNET financed under the Polish-Swiss Research Programme by Switzerland through the Swiss Contribution to the enlarged European Union (PSPB-146/2010, CARMNET), and by Poznan University of Technology under grant 04/45/DSPB/0122.

REFERENCES

[1] M. Glabowski, A. Szwabe, D. Gallucci, S. Vanini, and S. Giordano, "Cooperative internet access sharing in wireless mesh networks: Vision, implementation and experimentation of the CARMNET project," International Journal On Advances in Networks and Services, vol. 7, no. 1-2, 2014, pp. 25–36.

[2] A. Szwabe, P. Misiorek, and P. Walkowiak, "Delay-Aware NUM system for wireless multi-hop networks," in European Wireless 2011 (EW2011), Vienna, Austria, Apr. 2011, pp. 530–537.

[3] The CARMNET Project. [Online]. Available: http://www.carmnet.eu/ [retrieved: March, 2015]

[4] U. Akyol, M. Andrews, P. Gupta, J. D. Hobby, I. Saniee, and A. Stolyar, "Joint scheduling and congestion control in mobile ad-hoc networks," in The 27th IEEE International Conference on Computer Communications (INFOCOM 2008), Apr 2008, pp. 619–627.

[5] B. Radunović, C. Gkantsidis, D. Gunawardena, and P. Key, "Horizon: Balancing TCP over multiple paths in wireless mesh network," in Proceedings of the 14th ACM international conference on Mobile computing and networking, MobiCom 2008, 2008, pp. 247–258.

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

[7] R. Ramanathan and R. Rosales-Hain, "Topology control of multihop wireless networks using transmit power adjustment," in Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064), vol. 2, no. c. IEEE, 2000, pp. 404–413.

[8] R. Draves, J. Padhye, and B. Zill, "Routing in multi-radio, multi-hop wireless mesh networks," in Proceedings of the 10th annual international conference on Mobile computing and networking - MobiCom '04. New York, New York, USA: ACM Press, 2004, pp. 114–128.

[9] Y. Yang, J. Wang, and R. Kravets, "Interference-aware load balancing for multihop wireless networks," University of Illinois at Urbana-Champaign, 2005, pp. 1–16.

[10] W. Jiang, S. Liu, Y. Zhu, and Z. Zhang, "Optimizing Routing Metrics for Large-Scale Multi-Radio Mesh Networks," 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Sep. 2007, pp. 1550–1553.

[11] A. Subramanian, M. Buddhikot, and O. Miller, "Interference aware routing in multi-radio wireless mesh networks," 2006 2nd IEEE Workshop on Wireless Mesh Networks, 2006, pp. 55–63.

[12] P. Misiorek, P. Walkowiak, S. Karlik, and S. Vanini, "Sip-based aaa in delay-aware num-oriented wireless mesh networks," Image Processing and Communications, vol. 18, no. 4, 2013, pp. 45–58.

[13] P. Walkowiak, R. Szalski, S. Vanini, and A. Walt, "Integrating CARMNET system with public wireless networks," ICN 2014, The Thirteenth International Conference on Networks, February 2014, pp. 172–177.

[14] Malta NET. [Online]. Available: http://www.malta-net.pl [retrieved: March, 2015]

[15] M. Urbanski, M. Poszwa, P. Misiorek, and D. Gallucci, "Evaluation of the delay-aware num-driven framework in an internetwork environment," Journal of Telecommunications and Information Technology, no. 3, 2014, pp. 17–25.

[16] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling for maximum throughput in multihop radio networks," IEEE Transactions on Automatic Control, vol. 37, no. 12, Dec. 1992, pp. 1936–1949.

[17] Y. Yang and J. Wang, "Design Guidelines for Routing Metrics in Multihop Wireless Networks," 2008 IEEE INFOCOM - The 27th Conference on Computer Communications, Apr. 2008, pp. 1615–1623.

[18] M. Güneş, F. Juraschek, and B. Blywis, "An experiment description language for wireless network research," Journal of Internet Technology (JIT), Special Issue for Mobile Internet, vol. 11, no. 4, July 2010, pp. 465–471.

[19] J. Postel. Internet protocol. RFC0791, September 1981. [Online]. Available: http://tools.ietf.org/rfc/rfc0791.txt [retrieved: March, 2015]

[20] S. Ha, I. Rhee, and L. Xu, "CUBIC: A New TCP-Friendly High-Speed TCP Variant," ACM SIGOPS Operating Systems Review, vol. 42, no. 5, Jul. 2008, pp. 64–74.

 ISBN: 978-1-61208-398-8

Comparisons of SDN OpenFlow Controllers over EstiNet: Ryu vs. NOX

Shie-Yuan Wang
Department of Computer Science
National Chiao Tung University
Hsinchu, Taiwan
Email: shieyuan@cs.nctu.edu.tw

Hung-Wei Chiu
Department of Computer Science
National Chiao Tung University
Hsinchu, Taiwan
Email: hwchiu@cs.nctu.edu.tw

Chih-Liang Chou
EstiNet Technologies, Inc.
Hsinchu, Taiwan
Email: clchou@estinet.com

***Abstract*—SDN (Software-defined Networks) is a new approach to networking in which the control plane is extracted from the switch and put into the software application called the controller. In an SDN, the controller controls all networking switches and implements specific network protocols or functions. So far, the OpenFlow protocol is the most popular protocol used to exchange messages between the controller and OpenFlow switches. In this paper, we use the EstiNet OpenFlow network simulator and emulator to compare two open source popular OpenFlow controllers — Ryu and NOX. We studied the behavior of Ryu when it controls a network with loops and how quickly Ryu and NOX can find a new routing path for a greedy TCP flow after a link's status has changed. Our simulation results show that (1) Ryu results in the packet broadcast storm problem in a network with loops; (2) Ryu and NOX have different behavior and performance in detecting link failure and changing to a new routing path.**

Keywords–SDN; OpenFlow; Network Simulator.

I. INTRODUCTION

SDN (Software-Defined Networks) [1] is an emerging network architecture. It is a programmable, dynamic, adaptable, and well-managed network architecture. SDN extracts the control-plane functions from legacy switches and implements them into a software application called a controller. Nowadays, the OpenFlow [2] [3] protocol is the most popular protocol used for a controller to control SDN switches. Via this protocol, an OpenFlow switch learns the forwarding information from the controller and forwards incoming packets based on the received information. In an SDN network, people often call a network function implemented by a controller a "controller application." A controller application can implement a useful network function such as network virtualization. With these applications, network administrators can more easily manage an SDN network.

In an SDN network, the OpenFlow controller and its various applications work together to control the network and provide services. Before one introduces a new controller application into an SDN network, however, one must validate and evaluate its correctness, efficiency, and stability. There are several approaches that can be used for this purpose. One approach is creating an OpenFlow network testbed with real OpenFlow switches. Although the results of this approach are convincing, it incurs very high cost and the network settings cannot be very flexible. Another approach is via simulation in which all network switches, links, protocols, their operations, and the interactions between them are all simulated by a software program. Generally speaking, if the simulation modeling is correct enough, the simulation approach is a low-cost, flexible, scalable, and repeatable approach. This explains its wide uses in the research communities.

In this paper, we used the EstiNet OpenFlow network simulator and emulator [4] [5] to compare the behavior and performance of two popular OpenFlow controllers — Ryu [6] and NOX [7]. EstiNet uses an innovative simulation methodology called the ""kernel-reentering" simulation methodology to provide many unique advantages. When EstiNet simulates a network, each simulated host uses the (shared) real-world Linux operating system and allows any real-world Linux programs to run over it without any modification. For this property, the real-life Ryu and NOX controller programs can readily run over EstiNet to control many simulated OpenFlow switches and we can create simulation test cases to study the details of their behavior.

There are several other popular open source OpenFlow controllers [8] [9] [10]. The reason why we chose to study NOX and Ryu is because NOX is the world's first OpenFlow controller and Ryu is widely used with the OpenStack cloud operating system for cloud orchestration. Both Ryu and NOX are real-world applications written in the python language and they are runnable on any operating system supporting python. In this paper, we chose the learning bridge protocol (LBP) and spanning tree protocol (STP) controller applications to study. We observed the phenomenon when using Ryu as the OpenFlow controller in a network with loops and studied how quickly Ryu and NOX can find a new path for a greedy TCP flow after a link's status has changed. We also studied the impact of the address resolution protocol (ARP) on the path-changing time for a greedy TCP flow under the control of Ryu and NOX. Our results reported in this paper reveal the performance, behavior, and implementation flaws of NOX and Ryu over the tested network settings.

The rest of the paper is organized as follows. In Section II, we present some information about EstiNet OpenFlow network simulator and emulator to let readers know more about its special capabilities in conducting SDN researches. In Section III, we show the simulating settings used in this study. In Section IV, we explain the path-finding and packet-forwarding functions in Ryu and NOX. Performance evaluation results are presented in Section V. Finally, we conclude the paper in Section VI.

 ISBN: 978-1-61208-398-8

II. EstiNet OpenFlow Network Simulator and Emulator

The current version (as of the publication date) of EstiNet OpenFlow network simulator and emulator is 9.0. It can accurately simulate thousands of Ver 1.4.1 OpenFlow switches. It supports both of the simulation mode and the emulation mode. In the simulation mode, a real-world open source OpenFlow controller such as NOX, POX, Floodlight, or Ryu application program can directly run up on a controller node in the simulated network to control these simulated OpenFlow switches without any modification. In the emulation mode, these controller application programs can run up on an external machine that is different from the machine used to simulate OpenFlow switches to control these simulated OpenFlow switches. In additiion, in the emulation mode, if an OpenFlow controller has been implemented as a dedicated hardware device, it can remotely control the simulated OpenFlow switches in EstiNet via an Ethernet cable.

EstiNet, when running in the simulation mode, can accurately simulate the properties of the links that connect simulated OpenFlow switches. These properties include link bandwidth, link delay, link downtime, and the medium access control (MAC) protocol used over the link (e.g., IEEE 802.3 or IEEE 802.11, etc.). As a result, performance evaluation of traffic flows or the whole OpenFlow network can be accurately studied in EstiNet. Furthermore, since during simulation, the advancement of the simulation clock is accurately controlled by EstiNet, the performance simulation results of EstiNet are always realistic and repeatable, totally unaffected by the number of OpenFlow switches and hosts simulated by it. These unique and important capabilities enable us to conduct SDN research for various purposes. In this paper, we used these capabilities to compare the in-depth differences in the protocol operations of the NOX and Ryu SDN controllers.

III. Simulation Settings

Figure 1 shows the network topology that we used for this study. Each of node 3, 4, 5 and 11 simulates a host running the real-world Linux operating system (Fedora 14). On top of these nodes, any real-world Linux program can run without modification. Each of node 6, 7, 8, 9 and 10 simulates an OpenFlow switch supporting the OpenFlow protocol version 1.0. Node 1 is a simulated host on top which the Ryu or NOX OpenFlow controller program will run. Node 2 is a simulated legacy switch that connects all simulated OpenFlow switches together with the OpenFlow controller node. This formed network is the control-plane network. All TCP connections between simulated OpenFlow switches and the OpenFlow controller are set up over the control-plane network and the messages between Ryu/NOX and simulated OpenFlow switches are all exchanged over this control-plane network. In contrast, all simulated hosts, simulated OpenFlow switches and all links connecting them together form the data-plane network and the real applications running on simulated hosts will exchange their information over the data-plane network.

We studied two simulation cases on the network topology shown in Figure 1. We set the link delay and bandwidth to 10 ms and 100 Mbps for each link in these simulation cases. Both cases start at 0'th sec and ends at 120'th sec. In case 1, we only used Ryu as the OpenFlow controller. Because Ryu only uses LBP (Learning Bridge Protocol) to forward packets

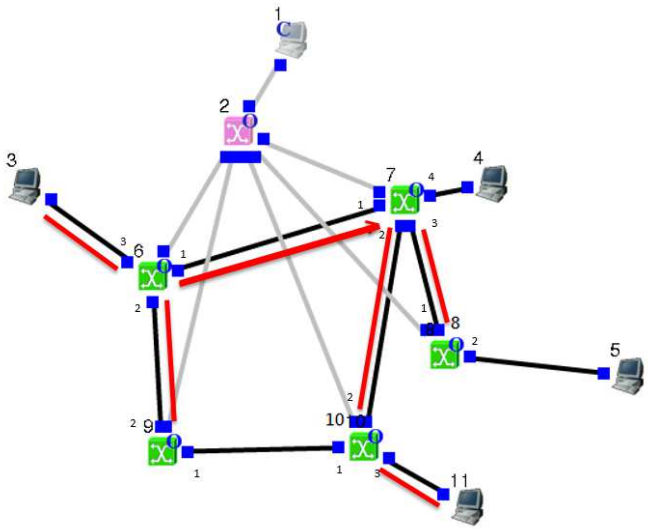


Figure 1. Spanning Tree in NOX before the link(6,7) breaks

without using SPT (Spanning Tree Protocol), we wanted to observe the phenomenon of running Ryu in a network with loops.

In case 2, we studied the required time to find a new path after a links status has changed when Ryu or NOX was used as the OpenFlow controller. We purposely broke the link between nodes 6 and 7 between 40'th sec and 80'th sec and shutdown the link between nodes 9 and 10 from 0'th sec to 40'th sec and from 80'th sec to 120'th sec. We studied the required time that the TCP flow can continue its transmission after a link on its path fails. We chose node 3 as the TCP sender and node 11 as the TCP receiver and generated endless TCP traffic from node 3 to node 11. Because over EstiNet one can directly run real-world Linux applications on simulated hosts without modification, we chose the open source programs "stcp" and "rtcp" as the TCP sender and TCP receiver programs. Once the stcp successfully sets up a real TCP connection with rtcp, it generates and sends endless TCP data to the rtcp. Both stcp and rtcp are set to start at 30'th sec rather than 0'th sec. This is because the OpenFlow controller needs enough time to discover the topology of the data-plane network and compute the path-finding and packet-forwarding information before any packet enters into the data-plane network.

We also studied the effects of the ARP protocol on the required time that the TCP flow finds a new path. Normally, on a real network the ARP protocol is enabled and the host will issue an ARP request to learn (MAC address, IP address) mapping information. However, avoid the ARP request/reply latency and bandwidth consumption, the ARP protocol can be disabled under some circumstances. Our simulation results show that when ARP is disabled, the path-finding capability and speed of NOX will significantly reduce.

IV. Path-Finding and Packet-Forwarding Functions in Ryu/NOX

In this section, we briefly explain how Ryu and NOX implement the path-finding and packet-forwarding functions.

 ISBN: 978-1-61208-398-8

A. LBP Component in Ryu

Ryu uses its LBP component to perform path-finding and packet-forwarding functions. When a switch issues a PacketIn message to Ryu due to a table-miss event, Ryu learns a "(switch ID, MAC address) = output port" mapping information from the packet that causes this table miss. This learned mapping information helps Ryu know through what "output port" the switch should forward a packet when its destination MAC address is the specified "MAC address" here. If Ryu currently has the mapping information for the destination host, it will send a FlowModify message to the switch to add a new flow entry and send a PacketOut message to the switch to forward the packet out of the specified port. Otherwise it will send a PacketOut message to the switch to flood the packet out of all of its ports. Here we use Figure 1 to explain how Ryu's LBP works. In order to explain how it works clearly, we assume that the link connecting nodes 9 and 10 is down on Figure 1 so that there is no loop in the network.

Suppose that the source host sends a TCP DATA packet to the destination host. When the packet arrives at node 6, because there are no flow entries in the flow table that can match it, node 6 sends a PacketIn message to Ryu and asks it how to process this packet. After Ryu receives the packet, it first learns (node 6, node 3's MAC) = port 3 mapping information. Because there are no mapping information about the destination host yet, it sends a PacketOut message to node 6 to flood the packet. After node 7 receives the packet, it issues a PacketIn message to Ryu and Ryu learns (node 7, node 3's MAC) = port 1. It then sends a PacketOut message to instruct node 7 to flood this packet. On the other hand, after node 9 receives the packet from node 6, the same scenario happens. It issues a PacketIn message to Ryu. Then, Ryu learns (node 9, node 3's MAC) = port 2 and sends a PacketOut message to node 9 asking it to flood this packet. However, because the link between nodes 9 and 10 is shutdown, node 10 cannot receive the packet from node 9. After node 10 receives the packet from node 7, the same scenario happens. Ryu learns (node 10, node 3's MAC) = port 2 and sends a PacketOut to instruct node 10 to flood this packet. When the destination host receives the packet, it sends a TCP ACK packet to the source host. When the packet arrives at node 10, because there are no flow entries that the packet can match, node 10 issues a PacketIn to Ryu and Ryu learns (node 10, node 11's MAC) = port 3. Now, with the mapping information learned before, Ryu issues a FlowModify message to node 10 instructing it to add a new flow entry of (destination MAC = host 3's MAC, ingress port = port 3, output port = port 2) and issues a PacketOut message to node 10 asking it to forward the packet out of port 2. After node 7 receives the packet, the same scenario happens. Ryu learns (node 7, node 11's MAC) = port 2. Then, Ryu issues a FlowModify message to node 7 instructing it to add a new flow entry of (destination MAC = host 3's MAC, ingress port = port 2, output port = port 1) and issues a PacketOut message to node 7 asking it to forward the packet to node 3 out of port 1. After node 3 receives the packet, the same scenario happens. Ryu sends a PacketOut message to node 3 asking it to forward the packet to node 3 out of port 3. After the TCP ACK packet enters node 3, the route from the destination host to the source host has completed and the related flow entries have been added into the switches. However, the route from the source host to the destination host is not completed yet. Later on, when node 3 sends the second TCP DATA packet to the destination host, after the packet enters node 6, the same scenario happens. Node 6 issues a PacketIn message to Ryu and Ryu issues a FlowModify message and PacketOut message to node 6. After the packet finally enters node 11, the route from the source host to the destination host finally is completed and the related flow entries have been added into the switches on the path. At this moment, the TCP flow can bidirectionally send its data along the path composed of nodes 6, 7, and 10 without bothering the Ryu controller.

B. STP and LBP Components in NOX

NOX's STP uses LLDP [11] packets to discover the topology of an OpenFlow network and build a spanning tree over the network. When an OpenFlow switch establishes a TCP connection to NOX, NOX immediately sends a FlowModify message to it and add a flow entry into its flow table. This flow entry will match future received LLDP packets and forward them to NOX. To discover the whole network topology, NOX sends LLDP packets to all switch ports in the network periodically. The LLDP transmission interval is 5 seconds. If there are N switch ports in the network, NOX will send a LLDP packet every (5 divided by N) seconds to evenly spread the LLDP traffic load. For each port of a switch, NOX will send a PacketOut message to the switch and ask it to send the LLDP packet carried in the PacketOut message out of the specified port every 5 seconds. Because NOX has taught the switch how to process LLDP packets before, when a switch receives a LLDP packet from other switches, it will send the received LLDP packet to NOX. With these received LLDP packets from switches, NOX knows the complete network topology and can build a spanning tree over it. For each link in the topology, NOX sends a PortModify message to the switches at the two endpoints of the link. This message sets the flooding status of the port connected to the link to FLOOD/NO_FLOOD according to whether the link is included/excluded in the spanning tree. NOX sets up a 10-second timer that is two times of the LLDP transmission interval to monitor a link's connectivity when it has been detected. When a link's timer expires, NOX thinks that this link is currently down and will build a new spanning tree. Then it sends a PortModify message to switches to change the flooding status of the affected ports.

NOX's LBP implementation is similar to Ryu's LBP implementation. The only difference is that NOX uses the spanning tree to prevent the packet broadcast storm problem. We use Figure 1 to explain how it works on this network topology. NOX uses STP to build a spanning tree as shown in Figure 1 and the spanning tree is composed of the links in red color and all links connecting a host to a switch . Suppose that the source node 3 sends a TCP DATA packet to the destination node 11. As discussed previously, when node 10 receives the flooded packet from node 7, it will issue a PacketIn message to NOX and then NOX issues a PacketOut message to node 10 to instruct it to flood the packet. However, because the status of port 1 of node 10 is set to NO_FLOOD, node 10 will not flood the packet on port 1.

V. Performance Evaluation

In this section, we studies two simulation cases. We first observed the phenomenon of Ryu when it operates in a network with loops. Then, we studied how quickly a TCP flow

 ISBN: 978-1-61208-398-8

can change its path to a new path under the control of Ryu and NOX when the topology changes.

A. Case 1

Our simulation results show that critical problems result when using Ryu as the OpenFlow controller in a topology with loops. These problems are the packet broadcast storm problem and insertion of incorrect flow entries into OpenFlow switches. We use Figure 1 to explain why these problems occur. Before the source host sends a TCP data packet to the destination host, it must broadcast the ARP request to learn the destination host's MAC address. After node 6 receives this ARP packet, it issues a PacketIn message to Ryu. Ryu learns the (node 6, node 3's MAC) = port 3 mapping information from this ARP packet but Ryu does not know any information about the destination MAC address, which is the broadcast address. As a result, Ryu sends a PacketOut message to node 6 asking it to flood the ARP packet. After node 7 and node 9 receive the packet, the same scenario happens and both nodes 7 and 9 flood the packet. Later on, node 10 receives the two ARP packets, one from node 7 and the other from node 9. Here we assume that the packet from node 9 arrives earlier than the packet from node 7. For each of these ARP packets, node 10 sends a PacketIn message to Ryu and Ryu sends a PacketOut message to node 10 asking it to flood the ARP packet.

When node 9 receives the ARP packet flooded from node 10, Ryu learns (node 9, node 3's MAC) = port 1 and overrides its information learned before. When node 7 receives the ARP packet flooded from node 10, Ryu learns (node 7, node 3's MAC) = port 2 and overrides its information learned before. When node 11 receives the ARP packet, it responds an ARP reply to node 3. When node 10 receives this ARP reply packet, node 10 issues a PacketIn message to Ryu and Ryu learns (node 10, node 11's MAC) = port 3. Ryu sends a FlowModify message to node 10 instructing it to add a new flow entry of (destination MAC = host 3's MAC, ingress port = port 3, output port = port 2) according to the information learned before. The idle timeout value and the hard timeout value associated with the flow entry are set to 0 by Ryu. Ryu then sends a PacketOut message to node 10 asking it to forward the packet out of port 2. This is because the ARP packet from node 7 was flooded on node 10 after the ARP packet from node 9 was flooded on node 10, which causes Ryu to learn that node 3 can be reached from port 2 of node 10.

According to the OpenFlow protocol, an idle timeout value specifies after how long the entry should be removed if no packet of this flow enters the switch to match this flow. On the other hand, a hard timeout value specifies after how long the entry should be removed after it has been added to the flow table. Since Ryu sets both the idle timeout and the hard timeout values to zero, it means that the flow entry is permanent and should never expire.

After node 7 receives the ARP reply packet, Ryu sends a FlowModify message to node 7 instructing it to add a new flow entry of (destination MAC = host 3's MAC, ingress port = port 2, output pot = port 2) and sends a PacketOut message to node 7 asking it to forward the packet out of port 2. (This is because node 7 received the ARP broadcast packet flooded from node 10, which made Ryu learn that node 3 can be reached from port 2 of node 7.) However, because the packet ingress port and output port are the same and the flow entry never expires, node 7 decides to drop the ARP reply packet. Because this incorrect flow entry has been added into node 7 and it will never expire due to the timeout value settings, from now on, any packet sent from node 11 to node 3 will be dropped at node 7. On the other hand, because Ryu cannot learn the forwarding information about the broadcast MAC address, it will ask OpenFlow switches to flood any received ARP request packet. As a result, for each ARP request packet received on any OpenFlow switch, it will be copied and flooded many times in the network. Worse yet, because there is no STP in the network, these ARP packets will spawn themselves repeatedly and eventually exhaust all the network bandwidth and cause the packet broadcast storm problem.

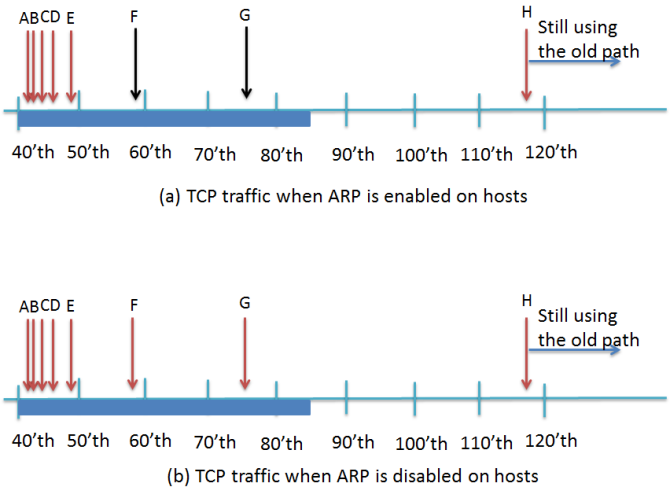


Figure 2. The timeline for a TCP flow to change/keep its path after the link between nodes 6 and 7 breaks at 40'th sec while using Ryu

B. Case 2 - Ryu

We studied how quickly the TCP flow can change its path to a new path under the control of Ryu. Our results show that the TCP flow never changes its path to a new path during the link(6,7) downtime and it becomes active again over the original path after the link downtime no matter whether ARP is enabled or disabled.

We use a timeline to display the significant events of TCP flow and ARP packet when ARP is enabled on hosts in Figure 2(a). Because we want to observe the related events of changing to the new path, our timeline only focuses on the interval from 40'th sec to 120'th sec. We denote the events A, B, C, D, E and H as the timestamps when the TCP flow tried to resend a lost packet due to the TCP reliable transmission design and the events F and G as the timestamps when the source host broadcasts the ARP request to learn the ARP record again. These ARP request transmissions occur because on the source host the ARP record for the destination host had expired during the long TCP retransmission interval. We found that there are two ARP packets but not TCP packets at the timestamps of events F and G. The H event represents the successful retransmission of the lost packet over the original path after the link between nodes 6 and 7 becomes up again at 80'th sec.

 ISBN: 978-1-61208-398-8

In the following, we explain (1) why the TCP flow cannot change to a new path during the link downtime and (2) why there are two ARP packets but no TCP packets at the timestamps of events F and G. The reason why the TCP retransmission fails at events A, B, C, D, and E are the same. Before the link breaks at 40'th sec, Ryu uses its LBP to find a path from the source host to the destination host, which traverses nodes 6, 7, and 10. Ryu also sends the FlowModify message and PacketOut message to these switches and set the idle timeout value and the hard timeout value associated with the flow entry to 0. At the timestamp of event A, when a TCP data packet enters node 6 after the link between nodes 6 and 7 breaks, because there is a flow entry in the flow table that the TCP data packet can match, node 6 forwards the packet out of port 1 to the broken link. Therefore, the destination host cannot receive any packets sent from the source host.

At the timestamp of event F, the source host broadcasts an ARP request before it retransmits the packet lost on the broken link. When the ARP request enters node 6, because there is no flow entry that can match a broadcast packet, node 6 sends a PacketIn message to Ryu. Ryu then sends a PacketOut message to node 6 asking it to flood the packet. As discussed previously, the ARP request will arrive at the destination host and the destination host will reply a unicast ARP reply packet to the source host. When the ARP reply enters node 11, node 11 will forward the packet out of port 2 to node 7 according to the flow entry that it learned before. When the ARP reply enters node 7, node 7 will forward the packet out of port 1 over the broken link. Therefore, the source host cannot receive the ARP reply, which made the source host unable to resend the TCP data packet. At the timestamp of event H, because the link downtime has passed, the source host can receive the ARP reply from the link between nodes 6 and 7 and successfully resend the TCP data packet. Finally, the TCP flow becomes active again over the original path.

When ARP is disabled on hosts, as shown in Figure 2(b), the TCP flow still cannot change to the new path during the link downtime. The reasons and the phenomenon are the same as when ARP is enabled on hosts, except for the events F and G. Because ARP is disabled on hosts, the source host uses the pre-built ARP table instead of broadcasting ARP request packets. Therefore, at the timestamps of events F and G, the source host tried to resend the TCP packet lost on the broken link rather than broadcasting the ARP request to learn the ARP record.

We found that the problem that a TCP flow cannot changes its path to a new path is caused by the improper settings of values of flow idle timeout and flow hard timeout. To fix this design flaw, we suggest to modify Ryu so that an installed flow entry can be expired after some idle period.

C. Case 2 - NOX

The NOX's spanning tree before the link(6, 7) fails is shown in Figure 1. After the link(6, 7) breaks at 40'th sec, NOX rebuilds the spanning tree (to save space, the new spanning tree is not shown in this paper). In the following, we show that (1) when ARP is enabled, the TCP flow will change its path to a new path traversing nodes 3, 6, 9, 10, and 11 at 58'th sec. However, the TCP flow never changes back to the original path after the link between nodes 6 and 7 becomes up again; and (2) when ARP is disabled. The TCP flow never changes its path to the new path when the link between nodes 6 and 7 was down from 40'th to 80'th and it becomes active again over the old path at 112.3'th sec. These results are caused by the settings of the idle timeout and hard timeout values used in flow entries and NOX's LBP and STP implementations.

We use the timeline in Figure 3(a) to display the significant events of a TCP flow when ARP is enabled on hosts. Because we want to observe the related events of changing to the new path, our timeline only focuses on the interval from 40'th sec to 120'th sec. Since the link between nodes 6 and 7 breaks at 40'th sec, as discussed previously, because NOXs STP monitors a link's status every 10 seconds, we expected to see the new spanning tree be built at around 50+'th sec and the TCP flow change to a new path after the new spanning tree is built. However, our simulation result shows that the TCP flow changes its path to the new path at around 59'th sec rather than at 50+'th sec.

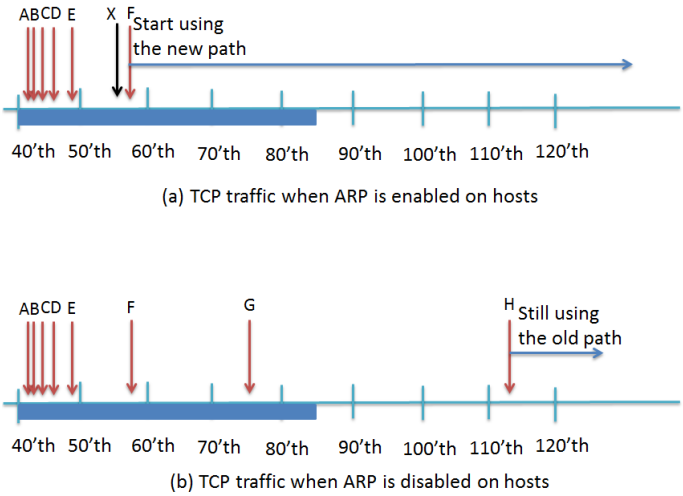


Figure 3. The timeline for a TCP flow to change/keep its path after the link between nodes 6 and 7 breaks at 40'th sec while using NOX

We denote the events A, B, C, D, E, and F as the timestamps when the TCP flow tried to resend a packet due to the TCP reliable transmission. The F event represents the successful retransmission of the lost packet over the new path during the link downtime. The X event represents the ARP request, which is triggered by the TCP flow retransmission at the timestamp of event F. Because the ARP record expires during the long TCP retransmission interval on the source host, the source host has to broadcast the ARP request to the network to learn the ARP record again.

In the following, we explain why the TCP flow changes to the new path at 50+'th sec when ARP is enabled. The reason why the TCP retransmission fails at events A, B, C, D, and E are the same and explained below. When NOX added a flow entry into a switch's flow table, the idle timeout value associated with the entry is set to 5 seconds. According to EstiNet and NOX logs, we found that NOX formed the new spanning tree at 51'th sec, which is after the timestamp of event E. Therefore, all of the resent TCP packets are forwarded over the broken link and got lost before the new spanning tree is formed. After NOX rebuilds the spanning tree, it sends

 ISBN: 978-1-61208-398-8

PortModify messages to switches instructing them to modify the statuses of their ports to FLOOD/NO_FLOOD according to the spanning tree's status.

As for the retransmission at event F, it succeeds and the reason is explained below. Because the interval between the timestamps of events E and F is larger than the value of flow entry's idle timeout, the flow entries on all switches will have expired by event F. Because the ARP record has expired as well on the source host, the source host broadcasts a ARP request at the timestamp of event X to learn the mapping information. After the ARP request enters node 6, because there is no flow entry in the flow table, node 6 sends a PacketIn message to NOX asking for forwarding instructions. NOX then sends a PacketOut message to node 6 asking it flood the ARP request out of all of its FLOOD ports. As a result, the ARP request traverses nodes 6, 9, and 10 to reach the destination host, and the destination host sends back the unicast ARP reply back to the source host.

We found that the ARP request and reply packets play very important roles. They not only let NOX learn the latest forwarding information but also install new and correct flow entries for ARP packets into all switches. Later on, when the resent TCP packet enters node 6, because there are no entries for this TCP flow in its flow table, node 6 sends a PacketIn message to NOX asking for instructions. NOX then sends a FlowModify message and PacketOut message to node 6 asking it to add a new flow entry for this TCP flow and forward the TCP packet out of port 2. After that, the following TCP DATA packets and their ACK packets follow the LBP scenario described before and starts to flow smoothly over the new path at 58'th sec.

In contrast to the fact that the TCP flow can change its path to the new path when ARP is enabled, we found that the TCP flow never changes its path to the new path when ARP is disabled, as shown in Figure 3(b). The reason why the TCP retransmission fails at events A, B, C, D, and E are the same as that described before. At the timestamp of event E, when the TCP DATA packet enters node 6, node 6 still forwards the packet out of port 1 due to the (old) matched flow entry. However, this entry has become incorrect now as the link between nodes 6 and 7 is already broken. At the timestamp of event F, because the interval between the timestamps of events E and F is larger than the value of the flow entry's idle timeout, each flow entry on every switches will have expired by event F. When the TCP DATA packet enters node 6 at event F, because the flow entry for this TCP flow has expired, node 6 sends a PacketIn message to NOX asking for instructions. However, because there were no broadcast ARP packets flooded over the network, NOX has no chance to update and correct its forwarding information. As a result, NOX sends a FlowModify message and PacketOut message to node 6 with incorrect forwarding information and instructs node 6 to forward the TCP packet over the broken link. Clearly, the packet cannot reach the destination host. For the retransmission event G, the same scenario occurs. Finally, at the timestamp of event H, because the link between nodes 6 and 7 has become up again at 80'th sec, the TCP flow uses its old path to successfully retransmit its lost packet and starts to flow smoothly.

Another significant problem we found in Figure 3(a) is that the TCP flow does not change its path from the new path to its original path after the link downtime, even though the spanning tree has been restored to the original one. This problem is caused by the improper settings of the idle timeout and hard timeout, which are set to 5 seconds and infinite, respectively. As long as the TCP flow sends some packets over the new path every 5 seconds, the flow entries in these switches will never expire. Since there are flow entries that can match the incoming TCP packet, these switches need not send PacketIn messages to NOX to ask for forwarding information. Therefore, NOX has no chance to install new flow entries into the flow tables of these switches and the TCP flow still uses the new path without changing back to its original path.

VI. Conclusion

In this paper, we used the EstiNet OpenFlow network simulator and emulator to compare the path-finding and packet-forwarding behavior of two widely used OpenFlow controllers — Ryu and NOX. NOX is chosen because it is the world's first OpenFlow controller; Ryu is chosen because it is widely used for cloud orchestration controller applications. Our simulation results show that Ryu lacks the spanning tree protocol implementation and will result in the packet broadcast storm problem when controlling a network with loops. When Ryu controls a network without a loop, after a link failure, we found that a TCP flow cannot change to a new path whether ARP is enabled or disabled. In contrast, we found that NOX enables a TCP flow to change to a new path when ARP is enabled but the TCP flow cannot change to a new path when ARP is disabled. As shown in our studies, the behavior of Ryu and NOX are very different. In the future, we plan to use EstiNet to compare more SDN controllers. Two other SDN controllers that are widely used are POX and Floodlight, each of which has a different implementation for path-finding and packet-forwarding functions. It is interesting to see how these four SDN controllers differ in these important functions.

References

[1] ONF, "Software-defined networking: The new norm for networks," Apr 2012, White Paper. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/ sdn-resources/white-papers/wp-sdn-newnorm.pdf

[2] N. Mckeown, T. Anderson, H. BalaKrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, J. Turner, Jennifer, Shenker, Scott, Turner, and Jonathan, "Openflow: enabling innovation in campus networks," ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, 2008, pp. 69–74.

[3] "Open Networking Foundation," 2012, URL: https://www.opennetworking.org/ [accessed: 2014-01-02].

[4] S.-Y. Wang, C.-L. Chou, and C.-M. Yang, "Estinet openflow network simulator and emulator," Communications Magazine, IEEE, vol. 51, no. 9, 2013, pp. 110–117.

[5] "EstiNet 8.0 OpenFlow Network Simulator and Emulator," URL: http://www.estinet.com [accessed: 2014-01-02].

[6] "Ryu OpenFlow Controller," URL: http://osrg.github.io/ryu/ [accessed: 2014-01-02].

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

[8] "Pox: A python-based openflow controller," URL: http://www.noxrepo.org/ [accessed: 2014-01-02].

[9] "Floodlight OpenFlow Controller," URL: http://www.projectfloodlight.org/floodlight/ [accessed: 2014-01-02].

[10] "OpenDaylight Controller," URL: http://www.opendaylight.org/ [accessed: 2014-01-02].

[11] "Link Layer Discovery Protocol," IEEE Standard 802.1 AB, 2009.

 ISBN: 978-1-61208-398-8

Network Partitioning Problem to Reduce Shared Information in OpenFlow Networks with Multiple Controllers

Hidenobu Aoki, Junichi Nagano, and Norihiko Shinomiya
Graduate School of Engineering
Soka University
Tokyo, Japan
Email: shinomi@ieee.org

***Abstract*—This paper proposes the layered control plane of OpenFlow networks with multiple controllers. Our method logically partitions an OpenFlow network and assigns controllers to the partitioned networks as their administrative domains. In addition, this paper focuses on the relationship between the network partitioning and the amount of global network information shared among controllers. Then, this paper handles the issue as a mathematical problem based on graph clustering and analyzes effective network partitioning methods in reducing the amount of global network information.**

Keywords*—*Software-Defined Networking; OpenFlow; multiple controllers; layered control plane; graph clustering.

I. INTRODUCTION

Software-Defined Networking (SDN) has been emerging as a new networking paradigm. The fundamental idea of SDN is to achieve programmable networking by separating the control and the data planes in an individual network device, such as a switch and router [1]. As one of the standard protocols between those separated planes, OpenFlow has been developed and widely used. In an OpenFlow network, a controller is in charge of generating data forwarding rules. In contrast, OpenFlow switches distributed in the data plane are responsible only for forwarding data according to the rules. This centralized architecture where a controller manages OpenFlow switches enables network operators to dynamically configure switches and to flexibly manage their networks [2].

Although it was assumed that a single controller dominates an entire network at the beginning, the concerns of scalability and reliability has been raised [3]. As the size of the network grows, the amount of data traffic, such as flow requests to a controller would increase [4] and [5]. Furthermore, the network operation with a single controller could take a risk of whole network breakdown if a failure occurs on the controller. As a result, to deploy multiple controllers has been considered.

In an OpenFlow network managed by multiple controllers, it could be scalable approach to logically divide the network into sub-networks as administrative domains of controllers so as to handle flow requests faster and reduce computational load on each controller. Then, a collaborative framework that enables controllers to effectively communicate each other has been required and drawn attention as a SDN-related research topic [6]. This paper proposes the layered architecture of control plane and classifies the roles of controllers. In particular, this paper focuses on how to decide the administrative domains of controllers, which has not discussed in any relevant work. The organization of this paper is as follows: In Section II, the related work of distributed controllers is addressed. Section III presents the layered control plane with its definitions and functions. In addition, the issue between the network partitioning and the amount of global network information shared among controllers is described. Section IV presents the graph definitions of the layered control plane and formulates the problem. As solutions of the network partitioning, Section V describes clustering algorithms. Section VI shows the simulation results. Finally, we conclude this paper with future work in Section VII.

II. RELATED WORK

Regarding to the deployment of multiple controllers, how to disseminate network state information over multiple controllers is highlighted as an important issue [4]. HyperFlow [7] realizes the synchronization of the network information among multiple controllers by utilizing a distributed file system called WheelFS WheelFS employs publish-subscriber patterns and contains all network information so that controllers can access to sufficient information for the local control of switches. In addition, ONOS [8] maintains a global network view with the abstraction of data plane network and shares topology information across ONOS servers by adopting distributed Titan graph database and Cassandra key-value store.

In order to reduce the load on controllers in sharing network information, a concept of layered control plane has been proposed. Onix [9] logically partitions a network, and controllers are assigned to partitioned networks as their control domains. Then, each partitioned network is contracted as a logical node and used as a unit for sharing network information among controllers. This enables a controller to communicate with other controllers without knowing specific network topology and states of other partitioned networks. In this way, the reduction of the amount of network information possessed by a single controller can be achieved.

Moreover, Kandoo [10] provides a layered control method for OpenFlow networks consisting of the root controller and some local controllers. The root controller manages all local controllers and is responsible only for the events which requires information over the whole network. On the other hand,

 ISBN: 978-1-61208-398-8

the local controller deals with the local events like requesting flow setups and collecting the statistics for a network with governing a group of switches and links between them. This layered control defines the scope of operations for processing different requests efficiently, which could offload the burden of the root controller.

As stated above, notable ideas of the layered control methods have been proposed in those related work. Nevertheless, they have not mentioned how to logically partition a network to decide administrative domains of controllers against switches although it could affect the amount of network information shared among controllers. In order to consider the issue, this paper focuses on the network partitioning and examines its solutions based on the concept of graph clustering.

III. Layered control plane

This section presents the architecture of the layered control plane and describes the issue between the network partitioning and the amount of global network information shared among controllers.

A. Definitions

In an OpenFlow network with multiple controllers, the network can be logically partitioned as administrative domains of controllers. In each domain, a controller is responsible for the following two roles: (1) management of switches in own domain and (2) federation of a whole network by communicating with other controllers. In accordance with those roles, the control plane can be layered in two tiers: the local tier and the federation tier in charge of (1) and (2), respectively.

B. Network Topology in Local and Federation Tiers

In the local tier, a local control function, called a local controller, describes the network topology of each administrative domain as a local graph. Moreover, the local graph is contracted to a single node which is used as a unit of communication with other controllers. In the federation tier, a global control function, referred to a federator, gathers the contracted nodes from all controllers and unifies them with edges between local graphs to form a federation graph describing global network topology. Because of this topology contraction, it is assumed to reduce the amount of global network information shared among controllers through the distributed database. Figure 1 illustrates an example of the layered control plane. In Figure 1, there are two domains described as local graphs 1 and 2. In the federation tier, on the other hand, the federation graph has two contracted nodes, and three edges correspond to the edges between the local graphs.

C. Network Topology Acquisition

In an OpenFlow network, the network topology is obtained by a combination of LLDP (Link Layer Discovery Protocol) packets, packet_in and packet_out messages as follows:

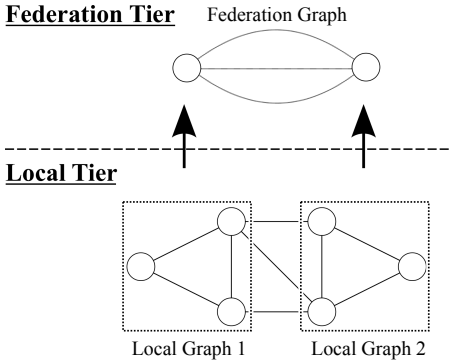


Fig. 1. An example of topology contraction in layered control plane.

1) Receiving a new packet, a switch forwards it to the controller as a packet_in message if it does not match any flow entries installed in the switch.
2) When the controller receives the packet_in message, the controller installs a flow entry to the switch if there exists the destination of the packet in its administrative domain; otherwise, in order to find the destination, the controller sends packet_out messages to direct switches to forward LLDP packets from its ports.
3) The switches send LLDP packets from their ports to adjacency switches. If the packet is received by a switch in the same domain, the packet_in message is sent to the controller; otherwise, it is forwarded to another controller.
4) When a controller receives the packet_in message, it inspects the message and detects the connectivity of switches.
5) If a controller detects the connectivity between switches within its domain, it updates own local database holding local network information; otherwise, as the controller detects an inter-domain link, it updates the distributed database keeping global network information.

Through LLDP flooding, controllers discover the destination of packets and install the corresponding flow entries in each switch in its administrative domain. Note that since LLDP packets will flood the entire network until the destination is found, this overhead will increase as the size of network becomes larger. Moreover, on the occasion of the inter-domain link detection, controllers need to update the distributed database, which may degrade their throughput. Furthermore, if there are many inter-domain links, controllers may need to access frequently to the distributed database to obtain and update the global network topology and compute routing paths. Therefore, we assume that the number of inter-domain links could affect the network performances and the load on controllers.

D. Network Partitioning in Local Tier

Considering the topology contraction in the layered control plane, it would be noteworthy that the number of inter-domain links, that is, the number of edges in a federation graph depends on how to partition a network; in other words,

 ISBN: 978-1-61208-398-8

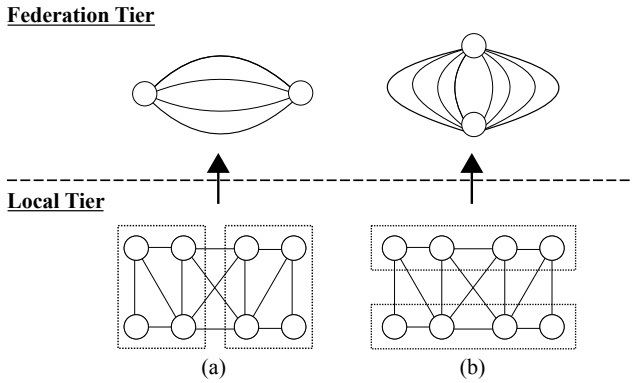


Fig. 2. An example of network partitioning in different ways.

how to decide administrative domains of controllers affects the amount of global topology information shared among controllers. Figure 2 shows an example of network partitioning in different ways. As seen in Figure 2, there are four edges in federation graph (a) while federation graph (b) has eight edges. This is because of different ways of network partitioning between (a) and (b). Therefore, it can be expected to further reduce the amount of global topology information by focusing on the network partitioning. This paper defines the problem to decide administrative domains of controllers as Network Partitioning Problem (NPP).

IV. Model Definitions and Problem Formulation

This section presents graph definitions treated in the layered control plane and formulates NPP.

A. Graph Definitions

Network topology can be described as a graph $G = (V, E)$: a set of vertices V represents network devices, such as routers and switches, and a set of edges E denotes links that connect those devices. In the local tier, local graphs are defined as

$$G_1^l = (V_1^l, E_1^l), \ldots, G_i^l = (V_i^l, E_i^l). \quad (1)$$

Note that a node can not belong to multiple local graphs. In the federation tier, on the other hand, a federation graph is denoted as

$$G^f = (V^f, E^f) \quad (2)$$

where G^f contains all local graphs as contracted nodes in V^f and edges between local graphs in E^f.

B. Problem Formulation as Graph Clustering

In the field of graph theory, graph clustering is defined as a task of grouping nodes in a graph into subsets called clusters in some predefined sense [11]. This paper applies the concept of graph clustering for network partitioning. Let a local graph G_i^l in (1) be a cluster, and a set of local graphs $\mathbf{G^l}$ is defined as a clustering: $\mathbf{G^l} = \{G_1^l, ..., G_i^l\}$. In general, it is regarded as a desirable clustering where there are many edges within each cluster called intra-cluster edges and relatively few edges between clusters referred to inter-cluster edges [12]. Considering the network topology treated in the layered control plane, intra-cluster edges correspond to edges in E_i^l, and inter-cluster edges are equivalent to edges in E^f. Consequently, we could say that the general criterion for the desirable clustering is applicable for the objective of NPP, that is, to partition a network such that the number of edges in a federation graph is reduced. Therefore, the objective function of NPP is defined as finding a clustering $\mathbf{G^l}$ such that

$$\text{Minimize } |E^f| \quad (3)$$

where an upper bound of the number of nodes in a cluster q is satisfied.

V. Clustering Algorithms for Network Partitioning

In this section, three clustering algorithms based on different measures are presented as solutions of NPP.

A. Minimum Cut Clustering

In graph theory, a minimum cut is defined as a set of the smallest number of edges which divide a graph into two disjoint subgraphs [13]. Based on the concept, we constructed Minimum cut clustering which separates a graph by minimum cut and regards the yielded subgraphs as clusters. It recursively conducts the separation process until the number of nodes in each cluster does not exceed the upper bound of the number of nodes in a cluster q as described in Algorithm 1.

Algorithm 1 Minimum Cut Clustering.

Require: $G = (V, E)$ and q: constraint of # of nodes
main
Clustering $\mathbf{G^l} \leftarrow \phi$
MinimumCutClustering(G, q)
Return $\mathbf{G^l}$
end main
function MinimumCutClustering(G, q)
(G_i^l, G_j^l) : *generate clusters based on minimum cut*
if # *of nodes in* $G_i^l \leq q$ **then**
  Add G_i^l to $\mathbf{G^l}$
else
  MinimumCutClustering(G_i^l, q)
end if
if # *of nodes in* $G_j^l \leq q$ **then**
  Add G_j^l to $\mathbf{G^l}$
else
  MinimumCutClustering(G_j^l, q)
end if
Return $\mathbf{G^l}$
end function

B. Conductance Clustering

As one of clustering indices, conductance has been defined, which compares the number of inter-cluster edges and that of intra-cluster edges yielded by a clustering [13]. By denoting a

set of all edges that have their origin in G_i^l and their destination in G_j^l as $E(G_i^l, G_j^l)$, the conductance of a cluster is defined as

$$\Phi(G_i^l) = \frac{|E(G_i^l, \mathbf{G^l} \setminus G_i^l)|}{\min(\sum_{v \in G_i^l} deg(v), \sum_{v \in \mathbf{G^l} \setminus G_i^l} deg(v))}. \quad (4)$$

Since finding a clustering with minimum conductance is known as NP-hard [12], we created Conductance clustering that chooses nodes one by one based on the conductance value as shown in Algorithm 2. The algorithm begins with a random node and assigns it to a cluster. Then, one of neighbor nodes of the node, which the cluster obtains the best conductance value, is chosen and assigned to the cluster. As this process, it expands the cluster by recursively choosing a neighbor node of the nodes in the cluster. If the number of nodes in the cluster reaches to the upper bound of the number of nodes in a cluster q, then it starts again to create a new cluster with a random node which has not belonged to any clusters.

Algorithm 2 Conductance Clustering.

Require: $G = (V, E)$ and q: constraint of # of nodes
Clustering $\mathbf{G^l} \leftarrow \phi$
$V' \leftarrow$ a list of nodes in a graph G
while $V' \neq \phi$ **do**
  $G_i^l \leftarrow \phi$
  Choose a node v_r randomly from V'
  Add v_r to G_i^l and remove v_r from V'
  while # *of nodes in* $G_i^l \leq q$ **do**
    for every neighbor node v_n of $\forall\ v \in G_i^l$ **do**
      if v_n is in V' **then**
        G_c^l = G_i^l
        Add v_n to G_c^l
        Calculate $\Phi(G_c^l)$
      end if
    end for
    Choose G_c^l with minimum conductance $\Phi(G_c^l)$
    Add v_n in the G_c^l to G_i^l
    Remove v_n from V'
  end while
  Add G_i^l to $\mathbf{G^l}$
end while
Return $\mathbf{G^l}$

C. Distance-k Cliques Clustering

In addition to minimum cut and conductance, distance is also a general clustering measure. We apply one of distance-based clustering algorithms called Distance-k cliques clustering [14]. Distance-k cliques clustering measures the strength of a relationship between two nodes in a graph in terms of the shortest path length between two nodes and generates clusters such that every pair of nodes is connected by a path of length at most k. As shown in Algorithm 3, Distance-k cliques clustering algorithm obtains an initial clustering at first. In an initial clustering, clusters are generated by choosing a node with the highest degree and its neighbors. Based on the initial clustering, the next step is to combine the clusters while both constraints of the number of nodes in a cluster q and the diameter of a cluster k are satisfied. Note that in order to fit the simulation setting of this paper, we added the constraint of the number of nodes in a cluster q which is not considered in the original algorithm defined in [14].

Algorithm 3 Distance-k Cliques Clustering.

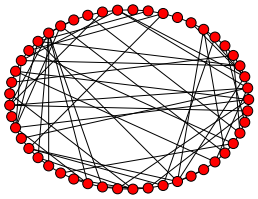
Require: $G = (V, E)$ and q: constraint of # of nodes
STEP1 : Obtain an initial clustering
Clustering $\mathbf{G^l} \leftarrow \phi$
$V' \leftarrow$ a list of nodes in a graph G
while $V' \neq \phi$ **do**
  $G_i^l \leftarrow \phi$
  Find the highest degree node v_{max} in V'
  while # *of nodes in a cluster* $\leq q$ **do**
    Add v_{max} to G_i^l and remove v_{max} from V'
    Add neighbor nodes of v_{max} to G_i^l and remove those nodes from V'
  end while
  Add G_i^l to $\mathbf{G^l}$
end while
STEP2 : Combine the clusters of the initial clustering
while True **do**
  Find u_{max}, a node connected with nodes in other clusters where the total # of nodes in adjacency clusters G_{adj}^l is the largest in $\mathbf{G^l}$.
  if # *of nodes in* $G_i^l(u_{max}) = q$ **then**
    Break
  end if
  for $\forall\ G_{adj}^l$ of $G_i^l(u_{max})$ **do**
    if # *of nodes in* $G_i^l(u_{max}) + G_{adj}^l < q$ **then**
      if $diameter(G_i^l(u_{max}) + G_{adj}^l) \leq k$ **then**
        Add nodes in G_{adj}^l to $G_i^l(u_{max})$
      end if
    end if
  end for
  Update $\mathbf{G^l}$ with $G_i^l(u_{max})$
end while
Return $\mathbf{G^l}$

VI. Simulation and Results

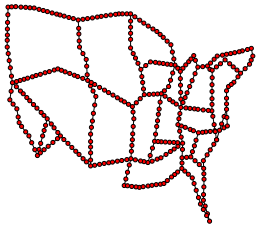
This section shows the simulation settings and results. We have developed a simulator in Python and NetworkX to evaluate three clustering algorithms in terms of minimizing $|E^f|$. In our simulation, those clustering algorithms are executed on a random graph called Newman Watts Strogatz (NWS) and two types of real network model, America and Japan models as shown in Figure 3. Note that a NWS graph is formed by connecting random pairs of nodes with a certain probability after creating a ring over n nodes [15]. The reason why we choose the random graph as well as real network models is to change the size of graph flexibly and examine the results. Therefore, we conduct two kinds of simulation for NWS graph: 1) varying the number of nodes and edges

 ISBN: 978-1-61208-398-8

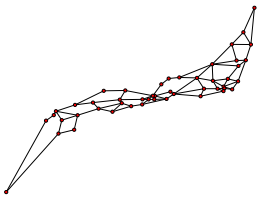
as described in Table I. 2) changing the value of the upper bound of the number of nodes in a cluster q. On the other hand, we test only 2) for the real network models since the size of the models is fixed as in Table I. Note that the distance constraint k of Distance-k cliques clustering is set to 10, and all simulations are executed 20 times.



(a) NWS [15].



(b) America Model [16].



(c) Japan Model [17].

Fig. 3. Examples of Network Model.

TABLE I. THE SIZE OF GRAPHS FOR SIMULATIONS.

	NWS			
The # of nodes	49	100	225	400
The # of edges	98	200	450	800
	America		Japan	
The # of nodes	365		48	
The # of edges	772		82	

A. Changing the Size of NWS Graph

Figure 4 shows $|E^f|$ on different number of nodes in the graphs where q is fixed to 15. The result indicates Conductance clustering demonstrates with the least $|E^f|$ on different size of the graph. In addition, as the size of the graph obtains larger, the difference of $|E^f|$ becomes considerable.

B. Varying the upper bound of nodes in a cluster

Figures 5 to 7 describe $|E^f|$ on different value of q from 10 to 25 while the number of nodes in a graph is fixed. Note that the number of nodes in a NWS graph is set to 225.

The negative slopes in those results indicate that as the value of q obtains greater, which means the larger number of nodes can be included in a cluster, it would yield the larger number of intra-cluster edges and the less number of inter-cluster edges. However, the results of Distance-k cliques clustering show horizontal slopes in Figures 5, 6, and a part of 7. This would be because of the constraint of distance k. Even if the number of nodes in a cluster does not reach to its limitation q, the distance constraint would be more influential because the algorithm does not allow a path length of any pairs of nodes to exceed k.

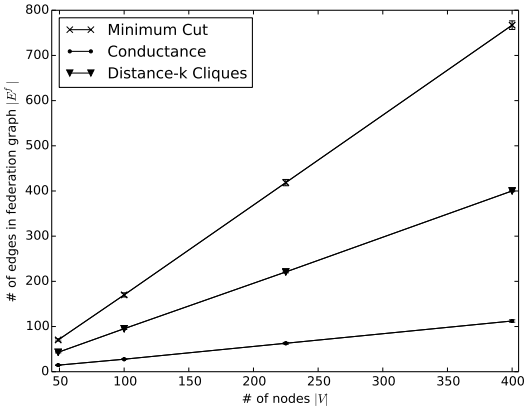


Fig. 4. $|E^f|$ on different size of NWS graph.

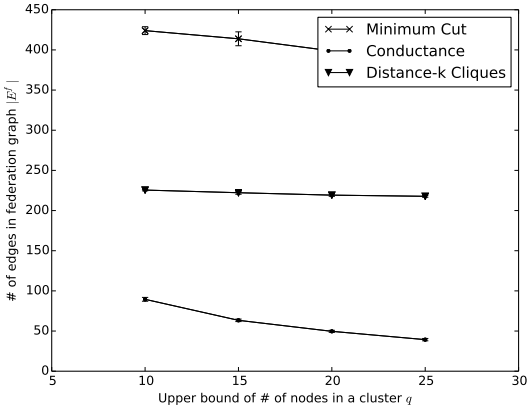


Fig. 5. $|E^f|$ on different value of q (NWS)

Furthermore, we can see that even when the value of q is varied, Conductance clustering reduces $|E^f|$ at most on any types of graph in our simulation (Figures 5 to 7). From the all results, we could say that conductance would be an important measure for effective network partitioning in reducing $|E^f|$, which indicates the reduction of the amount of shared topology information among controllers in OpenFlow networks. This is because Conductance clustering selects a node to assign a cluster based on the conductance value in (4) considering not only the number of cutting edges but also the density of yielded clusters. Therefore, it could have a tendency to yield a clustering with less $|E^f|$. On the other hand, Minimum cut clustering does not consider the quality of a clustering. It focuses on separating a graph based on the minimum number of cutting edges, which may results in separating only a small portion of a graph. As a result, the recursive graph separation by minimum cut could increase the number of clusters containing relatively small number of nodes, which ends up with a large number of inter-cluster edges as shown in our simulation results.

 ISBN: 978-1-61208-398-8

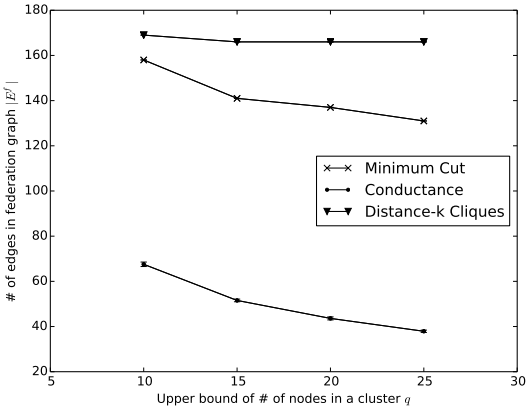


Fig. 6. $|E^f|$ on different value of q (America)

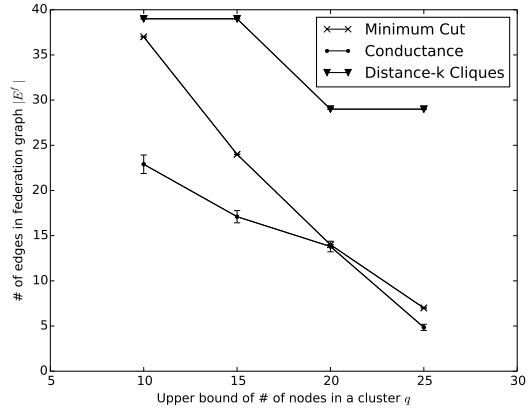


Fig. 7. $|E^f|$ on different value of q (Japan)

VII. Conclusion and Future Work

This paper proposed the layered control plane of OpenFlow networks and classified its functions in detail. In addition, the issue between the network partitioning and the amount of global network information in the layered control plane is addressed and converted to a mathematical problem as NPP. As a solution of NPP, this paper provided the network partitioning methods based on graph clustering and examined them on different network models. Our simulation results indicate that Conductance clustering performs the best in reducing the amount of network topology information shared among controller. Since our simulation is limited to theoretical approach, our future work will be an implementation of the layered control plane and examination of how reducing global network information by network partitioning can affect the load on controllers and the network performances under real network scenarios.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26330120.

References

[1] S. Sezer and et al., "Are we ready for SDN? implementation challenges for software-defined networks," Communications Magazine, IEEE, vol. 51, no. 7, July 2013, pp. 36–43.

[2] K. Suzuki and et al., "A survey on openflow technologies," IEICE Transactions on Communications, vol. 97, no. 2, 2014, pp. 375–386.

[3] S. Kuklinski and P. Chemouil, "Network management challenges in software-defined networks," IEICE Transactions on Communications, vol. 97, no. 1, 2014, pp. 2–9.

[4] S. Yeganeh, A. Tootoonchian, and Y. Ganjali, "On scalability of software-defined networking," Communications Magazine, IEEE, vol. 51, no. 2, February 2013, pp. 136–141.

[5] A. Tootoonchian, S. Gorbunov, Y. Ganjali, M. Casado, and R. Sherwood, "On controller performance in software-defined networks," in Proceedings of the 2nd USENIX conference on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services. USENIX Association, 2012, pp. 10–10.

[6] D. Marconett and S. Yoo, "Flowbroker: A software-defined network controller architecture for multi-domain brokering and reputation," Journal of Network and Systems Management, 2014, pp. 1–32.

[7] A. Tootoonchian and Y. Ganjali, "Hyperflow: A distributed control plane for openflow," in Proceedings of the 2010 Internet Network Management Conference on Research on Enterprise Networking, ser. INM/WREN'10, 2010, pp. 3–3.

[8] P. B and et al., "Onos: towards an open, distributed sdn os," in Proceedings of the third workshop on Hot topics in software defined networking. ACM, 2014, pp. 1–6.

[9] T. Koponen and et al., "Onix: A distributed control platform for large-scale production networks." in OSDI, vol. 10, 2010, pp. 1–6.

[10] S. Hassas Yeganeh and Y. Ganjali, "Kandoo: a framework for efficient and scalable offloading of control applications," in Proceedings of the first workshop on Hot topics in software defined networks. ACM, 2012, pp. 19–24.

[11] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," Journal of the ACM (JACM), vol. 51, no. 3, 2004, pp. 497–515.

[12] S. E. Schaeffer, "Graph clustering," Computer Science Review, vol. 1, no. 1, 2007, pp. 27–64.

[13] U. Brandes and T. Erlebach, "Network analysis." Springer Berlin Heidelberg, 2005.

[14] J. Edachery, A. Sen, and F. J. Brandenburg, "Graph clustering using distance-k cliques," in Graph drawing. Springer, 1999, pp. 98–106.

[15] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," Proceedings of the National Academy of Sciences, vol. 99, no. suppl 1, 2002, pp. 2566–2572.

[16] N. Shinomiya, T. Hoshida, Y. Akiyama, H. Nakashima, and T. Terahara, "Hybrid link/path-based design for translucent photonic network dimensioning," Journal of Lightwave Technology, vol. 25, no. 10, 2007, pp. 2931–2941.

[17] T. Sakano and et al., "A study on a photonic network model based on the regional characteristics of japan (in japanese," IEICE Technical Report, PN2013-01, Tech. Rep., 2013.

 ISBN: 978-1-61208-398-8

State-of-the-art Energy Efficiency Approaches in Software Defined Networking

Beakal Gizachew Assefa and Oznur Ozkasap
Department of Computer Engineering
Koc University, Istanbul, Turkey
e-mail: [bassefa13, oozkasap]@ku.edu.tr

***Abstract*—Software Defined Networking (SDN) paradigm has been attracting an increasing research interest. Promising features of SDN are enabling programmable network components and separating the control plane and the forwarding plane. It offers several advantages such as flexibility without sacrificing forwarding performance, high efficiency through optimized routing, ease of implementation and administration, and cost reduction. On the other hand, energy efficiency in networking is an issue as energy cost contributes significantly to the overall costs in information and communication technologies. Thus, energy efficient mechanisms for SDN components have become indispensable. In this study, we address the importance of energy efficiency mechanisms in SDN, propose a classification of methods for improving energy efficiency in SDN and describe each group of solutions with their principles, benefits and drawbacks. To the best of our knowledge, our study is the first one that focuses on state-of-the-art energy efficiency strategies for SDN, discusses open issues and provides guidelines for future research.**

Keywords–Software Defined Networking; SDN; Energy efficiency

I. INTRODUCTION

Software Defined Networking (SDN) is a recent trend in computer networks based on the concepts of control plane and data (forwarding) plane separation and logically centralized control by enabling programmable network devices [1][2]. The idea behind SDN paradigm, depicted in Fig.1, is to eliminate the tight coupling between control and forwarding components in traditional network design, and hence the drawbacks of cumbersome network configuration and limited flexibility to changing requirements. A logically centralized controller configures the forwarding tables (also called flow tables) of switches, which are responsible for forwarding the packets of communication flows.

SDN has been deployed in a diverse set of platforms, ranging from home networks and institutional networks to data center networks. It promises several advantages such as flexibility without sacrificing forwarding performance, high efficiency through optimized routing, ease of implementation and administration, and cost reduction.

The energy consumption constitutes a significant portion of overall information and communication technology costs [3][4]. Several research studies have been conducted for reducing energy costs in different network settings such as wireless sensor networks [5] and cloud data centers [6]. However, to the best of our knowledge, a survey on recent energy saving strategies for SDNs is not available. In this study, we propose a classification of methods for improving energy efficiency in SDN and discuss each group of solutions. We also identify open issues and future research directions.

Energy optimization can be applied at various components of the SDN architecture or SDN itself can be used as a means of energy saving. Energy saving in SDN can be addressed algorithmically or through hardware-based improvements. The hardware-based solutions are applied on the forwarding switches, and such solutions range from compressing the content of Ternary Content Addressable Memory (TCAM) to increasing the capacity of TCAM. Software-based solutions are applied on the controller. We classify state-of-the-art energy efficiency strategies for SDN into four categories, namely traffic aware, compacting TCAM, rule placement, and end host aware.

Traffic aware energy efficiency approaches are inspired by the fact that network components are often under utilized. The key principle is to turn on or off network components (i.e., SDN forwarding switches) based on the traffic load. For instance, when the traffic load is low (e.g., during night time) this approach has the potential to save up to 50% of the total energy consumption [7].

Typically, an elastic tree structure is used to represent the network components that can grow and shrink with the dynamic traffic load. The key challenge is to determine which components to turn off and turn on without compromising the required quality of service (QoS) [7]–[11].

Compacting TCAM solutions attempt to minimize the memory need of information stored in forwarding switches. In SDN, forwarding switches use TCAM, which is a specialized type of high-speed memory that performs an entire memory search in a single clock cycle. However, TCAM is very expensive and power hungry. A memory optimal strategy can be achieved by compacting TCAM itself or compressing the information stored in TCAM [12]–[15].

Rule placement techniques focus on how to place the rules in the forwarding switches. Given the network policies and end point policies, the controller provides a way to convert the high level policies into switch understandable rules. Rule placement is an NP-hard problem which needs a heuristic based solution. Although heuristic based approaches do not guarantee optimal solutions, they typically offer close to optimal results depending on the constraints [16][17][18]. Some of the constraints presented in this area are the maximum number of rules a switch can hold, the routing policy, and the topology. Under such constrains, these approaches attempt to optimize routing.

 ISBN: 978-1-61208-398-8

End host aware energy saving solutions use the practice of turning off underutilized physical servers and running their tasks on a fewer number of servers in SDN based data centers [19][20]. Specifically in data centers, the SDN model is used to form an overlay connecting virtual machines. Server virtualization helps systems to run multiple operating systems and services on a single machine.

The rest of the paper is organized as follows. Section II presents our classification and details of energy efficiency techniques for SDN. In Section III, we discuss open issues on energy efficiency in SDN and provide guidelines for future research. Finally, we give concluding remarks in Section IV.

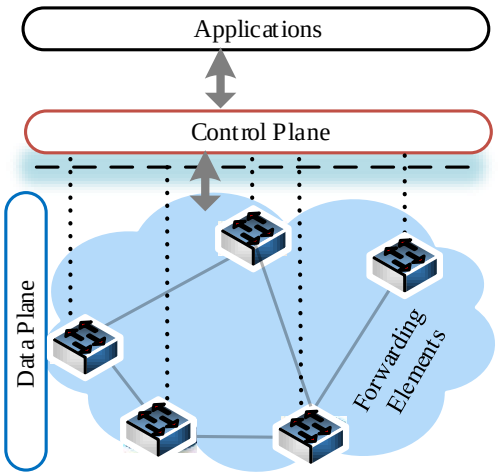


Figure 1: Software Defined Networking Architecture

II. Classification of Energy Efficient SDN Solutions

In this section, we describe solutions for improving energy efficiency in SDN, based on our classification of techniques, namely traffic aware, compacting TCAM, rule placement, and end host aware. Table I presents a summary of methods under each category along with their key properties.

A. *Traffic Aware Solutions*

With traffic aware energy efficiency approaches, energy consumption can be reduced by turning off some forwarding switches during low traffic load, or putting CPUs or ports at sleep mode. The solutions in this group have the potential to significantly improve energy efficiency in SDN. For data centers, traffic aware approach achieves power savings of up to 50% during low load periods [7].

ElasticTree is a power management solution for data center networks which is implemented on a testbed consisting of OpenFlow switches [7]. The idea is to turn off links and switches based on the amount of traffic load. Therefore, energy consumption of the network is proportional to the dynamically changing traffic.

ElasticTree consists of three optimizers: formal model, greedy bin-packing, and a topology-aware heuristic. Each optimizer takes network topology (a graph), routing constraints, power model (flat), and traffic matrix as input, and outputs subset of links and flow routes.

The formal model formulates the power saving problem by specifying objective function and constraints. The objective function minimizes the sum of the total number of switches turned on and the number of links. The advantage of the formal model is that it guarantees a solution within some configurable optimum; however, it only scales up to 1000 hosts.

The greedy bin-packing optimizer evaluates possible paths and chooses in left-to-right order manner that is the left most path is chosen first. The optimizer improves the scalability of the formal model. This approach suffers the same problem as any of heuristic techniques. However, solutions can be computed incrementally and can support on-line usage.

The topology-aware heuristic optimizer, on the other hand, splits the flow and finds the link subset easily. It is computationally efficient, since it takes advantage of a fat tree structure and takes only port counters to compute link subset. This approach uses IP to formalize the optimization problem. The drawback is degradation of performance because of turning on and turning off components.

CoRrelation-aware Power Optimization(CARPO) algorithm dynamically consolidates traffic flows onto a small set of links and switches in a data center network, and then shuts down unused network devices for energy savings [8]. It consolidates traffic flows based on correlation analysis among flows. Another important feature of CARPO is to integrate correlation-aware traffic consolidation with link rate adaptation for maximized energy savings. The integration is formulated as an optimal flow assignment problem. A near-optimal solution is first computed using linear programming to determine consolidation and the data rate of each link in the data center network. A heuristic algorithm is used to find a consolidation and rate configuration solution with acceptable runtime overheads. The heuristic reduces the computation complexity.

REsPoNse is a framework that allows network operators to automatically identify energy-critical paths [9]. It investigates the possibility to pre-compute a few energy-critical paths that, when used in an energy-aware fashion, can continuously produce close-to-optimal energy savings over long periods of time. REsPoNse identifies energy-critical paths by analyzing the traffic matrices, installs them into a small number of routing tables (called always-on, on-demand, and fail-over), and uses a simple, scalable online traffic engineering mechanism to deactivate and activate network elements on demand. The network operators can use REsPoNse to overcome power delivery limits by provisioning power and cooling of their network equipment for the typical, low to medium level of traffic.

A similar technique named Carrier Grade is proposed in [10]. While the Openflow architecture may not be able to significantly reduce energy consumption by consolidating the control hardware/software in a single machine, it shows significant promise by facilitating network wide energy efficiency solutions. MLTE was implemented in combination with local

 ISBN: 978-1-61208-398-8

TABLE I: SUMMARY OF ENERGY EFFICIENCY TECHNIQUES IN SDN

Category	Approach	Properties
Traffic aware	ElasticTree [7]	ElasticTree, based # traffic, Mixed IP, re-computation cost
	CARPO [8]	Fat Tree, based on correlation analysis among flows
	REsPoNse [9]	FatTree, identify energy-critical path to optimize
	Carrier Grade [10]	MLTE implementation, topology-aware heuristic
	Integrated [11]	Combined sleep and turning off, recovery from failure
Compacting TCAM	Rectilinear [12]	Rectangle Rule List (RRL) minimization, geometric model
	TCAM Razor [13]	Decision diagrams, dynamic programming, and redundancy removal
	Bit Weaving [14]	Non-prefix based compression
	Compact TCAM [15]	Usage of short tag
Rule Placement	Big Switch [17]	Heuristics for endpoint policy, routing policy and rule placement
	Palette [16]	Graphs, algorithms, heuristics
	Optimizing Rule Placement [18]	Meaning of rules, integer linear programming formulation of the problem
End host aware	Honeyguide [19]	VM migration, fault tolerance, easy deployment
	EQVMP [20]	Virtual Machine, load balancing

power saving options such as controlled adaptive line rates in the Openflow switches [21]. The technique is also extended to handle failures which happen in the controller or forwarding switches.

An integrated scheme that combines smart sleeping and power scaling based on the topology-aware heuristic algorithm to improve energy saving level of data center networks is proposed in [11]. This combined mechanism was deployed in a data center using Fat-Tree topology, and the bounds on energy savings in low and high traffic utilization cases were analysed. Analytical results show that the combined algorithm reduces energy consumption remarkably compared to the conventional one in case of high traffic.

B. Compacting TCAM Solutions

Content Addressable Memory (CAM) is a special type of memory that enables direct query to the content without having to specify its address. A search in CAM provides a search tag which is the representation or the content itself, and returns the address of the content if the item is found. The content is represented in binary (i.e., 0 and 1). TCAM is a specialized CAM, and the term ternary refers to the memory's ability to store and query data using three different inputs: 0, 1 and X. The X input, which is often referred to as a wildcard state, enables TCAM to perform broader searches based on pattern matching. TCAM is popular in SDN switches for fast routing lookup and it is much faster than RAM. However, TCAM is expensive, consumes high amount of power, and generates a high level of heat. For example, a 1Mb TCAM chip consumes 15-30 watts of power, and TCAM is at least as power hungry as SDRAM. Power consumption together with the consequent heat generation is a serious problem for core routers and other networking devices [13][22].

Two kinds of compression that can be applied in TCAM are rule compression and content compression. In a traditional Access Control List, a rule has five components: source range, destination range, protocol, port(s), and action. In SDN, the forwarding decision of a switch is based on flow tables implemented in TCAM. Each entry in the flow table defines a matching rule and is associated with an action. Upon receiving a packet, a switch identifies the highest-priority rule with a matching predicate, and performs the corresponding action. The proposals in [12]-[15] are attempts to compact these rules to utilize TCAM effectively.

Rectilinear [12] is an approach that exploits SDNs features such as programming interface to the switches and dynamic determination of actions for each flow at the switches. The compacting reduces the size of bits to store information that are essential to classify packets to a flow. A flow-id is given to each flow to uniquely identify packets in the corresponding flow. The packet headers are modified at the forwarding switches to carry the flow-id that can be used by other switches on the path for classifying the packets. A shorter tag representation for identifying flows than the original number of bits are used to store the flow entries of SDN switches. The authors demonstrated that the compact representation on flow can reduce 80% TCAM power consumption on average.

TCAM Razor proposes a four step solution to compress the packet classifier [13]. First, it converts a given packet classifier to a reduced decision diagram. Second, for every non-terminal node in the decision diagram, it minimize the number of prefixes associated with its outgoing edges using dynamic programming. Third, it generates rules from the decision diagram. Last, it removes redundant rules.

The Bit Weaving technique employs a non-prefix compression scheme [14], and is based on the observation that TCAM entries that have the same decision but whose predicates differ by only one bit can be merged into one entry by replacing the bit in question with *. Bit weaving consists of two new techniques, bit swapping and bit merging. First it swaps bits to make the similar and then merge such rules together. The key advantages of bit weaving are that it runs fast, it is effective, and can be complementary to other TCAM optimization methods as a pre/post-processing routine.

A Compact TCAM approach that reduces the size of the flow entries is proposed in [15] that uses shorter tags for identifying flows than the original number of bits used to store

 ISBN: 978-1-61208-398-8

the flow entries for SDN switches. The catch for this approach comes from the dynamic programming capability of SDN to route the packets using these tags. Furthermore,the usage of SDN framework to optimize the TCAM space is introduced.

C. Rule Placement Solutions

The energy efficiency techniques for rule placement solutions start by formalizing the energy cost model and the constraints associated, then applies heuristic technique to find optimum energy saving strategy. Forwarding rules are generated and pushed to the forwarding switches by the controller. The controller generates rules and pushes them to the forwarding switches. Placing the rules to respective switches distributed across the network and optimizing given an objective function under the constrains is NP-hard problem. The objective function in our particular case is minimizing energy where as the constrains are the the number of switches, flow table capacity, link capacity, and number of ports per switch. SDN makes use of a logically centralized controller which has a global view of the network that provides flexibility for optimizing forwarding routes.

The Palette distribution framework is a distributed approach applied to SDN tables [16]. Since the SDN controller table can only handle hundreds of entries, and the memory is expensive and power hungry, Palette decomposes large SDN tables into small ones and then distributes them across the network, while preserving the overall SDN policy semantics. Palette helps balance the sizes of the tables across the network, as well as reduce the total number of entries by sharing resources among different connections. It copes with two NP-hard optimization problems: decomposing a large SDN table into equivalent sub-tables, and distributing the sub-tables. The problem of traversing is formulated using the rain-bow problem.By giving unique color for each sub-tree, each connection traverses each color type at least once. Implementation of Palette is based on graph theory formulation algorithms and heuristics.

Big Switch approach utilizes the fact that SDN controllers have a global view of the network and proposes that the entire network should be viewed as one big switch [17]. The architecture modularizes the SDN controller into three components: endpoint policy, routing policy and rule placement policy. A high-level SDN application defines end-point connectivity policy on top of big switch abstraction; a mid level SDN infrastructure layer should decide on the routing policy; and a compiler should synthesize on the end-point and routing policy and develop an effective set of forwarding rules that obey the user-defined policies and adhere to the resource constraints of the underlying hardware. Minimizing the number of rules needed to realize the endpoint policy under rule capacity constraint is both a decision and an optimization problem. The architecture addresses the two problems through a heuristic algorithm that recursively covers the rules and packs groups of rules into switches along the path.

The drawbacks of Pallete and Big Switch approaches are that they do not rely on the exact meaning of the rules and the rules should not determine the routing of the packets.A technique of compacting rules, which enhances the rule placement, is proposed in [18]. This approach analizes the meaning of the rules, and together with heuristic optimization method to minimize energy consumption for a backbone network while respecting capacity constraints on links and rule space constraints on routers. They present an exact formulation using integer linear programming, and introduce efficient greedy heuristic algorithm for large networks.

D. End Host Aware Solutions

Server virtualization enables the working of multiple virtual machines simultaneously on a single physical server, thus decreasing electricity consumption and wasting heat as compared to running underutilized physical servers. Hence, instead of operating many servers at low utilization, virtualization technique combines the processing power onto fewer servers that operate at a higher total utilization. The deployment of SDN in cloud data center virtual machines boosts QoS and load balancing. Unlike traffic aware strategies, where the network components are the focus for energy saving, the virtual machines are utilized in the saving strategy.

Honeyguide is a virtual machine migration-aware network topology for energy efficiency in data center networks [19]. Reducing energy consumption is achieved by decreasing the number of active (turned on) networking switches. In this approach, the focus is not only turning off inactive switches, but also trying to maximize the number of inactive switches. To increase the number of inactive switches, two techniques are combined: virtual machine (VM) and traffic consolidation. As an extension of existing tree based topologies, Honeyguide adds bypass links between the upper-tier switches and physical machines. By doing so, it meets the fault tolerance requirement of data centers. It is easily deployable since what it needs is to add a bypass link only.

EQVMP proposes energy-efficient and QoS-aware virtual machine placement for SDN based data centers [20]. Unlike ElasticTree, power on and off is applied to the servers themselves. EQVMP combines three techniques: hop reduction, energy saving and load balancing. Hop reduction divides the VMs into groups and reduces the traffic load among groups by graph partitioning. Energy savings mostly are achieved by VM placement. The motivation behind VM placement is from Best Fit Decreasing (BFD) and Max-Min Multidimensional Stochastic Bin Packing. Fat-tree is used to represent the VM and servers in the data center. SDN controller is used to balance the load in the network. Load balancing achieves flow transmission in networks without congestion.

III. Open Issues and Discussion

For energy efficiency techniques in SDN, we identify the following open issues and future research directions.

- In traffic aware solutions, turning the network components on and off based on network load helps in reducing energy consumption. However, determining the set of network components to turn on or turn off dynamically without affecting QoS and performance is an NP-hard problem. An efficient solution in this area should consider the trade off between energy savings and network performance.
- Other open research issues in traffic aware energy efficiency are scalability and flexibility. The traffic aware

 ISBN: 978-1-61208-398-8

model needs to be scalable in the case of high traffic load. Flexibility is the ability of the system to adapt dynamic network settings (i.e., different topologies, change in the number of nodes).

- Forwarding switches use TCAM, which is expensive and power hungry. Some research proposals have attempted to compact the rules stored in TCAM. For such compacting TCAM solutions, information stored in TCAM cannot be further compressed after a certain threshold.
- Formal definition of the energy saving problem is the base for applying a sound theoretical solutions. Since the problem is NP-hard, utilizing heuristic techniques is inevitable. On an average case, heuristics can give close to optimal solution with in a feasible time. Specially, in a dynamic environment, the problem becomes even more challenging.
- Rule placement directly affects both the performance of the network and also determines the routing. Given a routing policy and end policy of a network, there is a need for formal and more space efficient way of representing rules.
- There have been few studies in end host aware energy efficiency strategies. Combining the advantages of virtualization with SDN can improve performance and minimize energy consumption.

IV. Conclusion

In this paper, we address the importance of energy efficiency mechanisms in software defined networks. With the significance of energy efficiency in networking, mechanisms enabling energy savings in the SDN model become indispensable. We propose a classification of methods for improving energy efficiency in SDN (traffic aware, compacting TCAM, rule placement, and end host aware) and discuss each group of solutions. To the best of our knowledge, our study is the first one that focuses on recent energy saving strategies for SDN. As future work, we aim to conduct a comprehensive study on the methods, make a experimental comparisons, and develop measurement metrics.

Acknowledgements

This work was partially supported by the COST (European Cooperation in Science and Technology) framework, under Action IC0804 (Energy Efficiency in Large Scale Distributed Systems), and by TUBITAK (The Scientific and Technical Research Council of Turkey) under Grant 109M761.

References

[1] B. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," Communications Surveys Tutorials, IEEE, vol. 16, no. 3, 2014, pp. 1617–1634.

[2] K. Diego et al., "Software-Defined Networking: A Comprehensive Survey," Proceedings of the IEEE, vol. 103, no. 1, Jan 2015, pp. 14–76.

[3] L. Chiaraviglio, M. Mellia, and F. Neri, "Minimizing isp network energy cost: Formulation and solutions," IEEE/ACM Transactions on Networking, vol. 20, no. 2, April 2012, pp. 463–476.

[4] M. Pickavet et al., "Worldwide energy needs for ict: The rise of power-aware networking," in Advanced Networks and Telecommunication Systems (ANTS), 2008. 2nd International Symposium on, Dec, pp. 1–3.

[5] R. Soua and P. Minet, "A survey on energy efficient techniques in wireless sensor networks," in Wireless and Mobile Networking Conference (WMNC), 2011, 4th Joint IFIP, Oct, pp. 1–9.

[6] A. Hammadi and L. Mhamdi, "A survey on architectures and energy efficiency in data center networks," Computer Communications, vol. 40, 2014, pp. 1 – 21.

[7] B. Heller et al., "Elastictree: Saving energy in data center networks," in NSDI, 2010.

[8] X. Wang, Y. Yao, X. Wang, K. Lu, and Q. Cao, "Carpo: Correlation-AwaRe power optimization in data center networks," in IEEE INFOCOM, March 25-30, 2012, pp. 1125–1133.

[9] N. Vasić et al., "Identifying and using energy-critical paths," in Seventh Conference on Emerging Networking EXperiments and Technologies, ser. CoNEXT, 2011. ACM, pp. 18:1–18:12.

[10] D. Staessens, S. Sharma, D. Colle, M. Pickavet, and P. Demeester, "Software defined networking: Meeting carrier grade requirements," in Local Metropolitan Area Networks (LANMAN), 2011 18th IEEE Workshop on, Oct, pp. 1–6.

[11] T. Nguyen et al., "Modeling and experimenting combined Smart sleep and power scaling algorithms in energy-aware data center networks," Simulation Modelling Practice and Theory, vol. 39, 2013, pp. 20 – 40.

[12] D. A. Applegate et al., "Compressing rectilinear pictures and minimizing access control lists," in ACM-SIAM Symposium on Discrete Algorithms (SODA), 2007, pp. 1066–1075.

[13] C. R. Meiners, A. X. Liu, and E. Torng, "Tcam razor: A systematic approach towards minimizing packet classifiers in tcams," in 15th IEEE International Conference on Network Protocols (ICNP), Beijing, China, October 2007.

[14] C. Meiners, A. Liu, and E. Torng, "Bit weaving: A non-prefix approach to compressing packet classifiers in tcams," IEEE/ACM Transactions on Networking, vol. 20, no. 2, April 2012, pp. 488–500.

[15] K. Kannan and S. Banerjee, "Compact TCAM: flow entry compaction in TCAM for power aware SDN," IEEE/ACM Transactions on Networking, vol. 20, no. 2, Apr. 2012, pp. 488–500.

[16] Y. Kanizo, D. Hay, and I. Keslassy, "Palette: Distributing tables in software-defined networks," in IEEE INFOCOM, April 2013, pp. 545–549.

[17] N. Kang, Z. Liu, J. Rexford, and D. Walker, "Optimizing the "one big switch" abstraction in software-defined networks," in Ninth ACM Conference on Emerging Networking Experiments and Technologies, ser. CoNEXT. New York, NY, USA: ACM, 2013, pp. 13–24.

[18] F. Giroire, J. Moulierac, and T. K. Phan, "Optimizing Rule Placement in Software-Defined Networks for Energy-aware Routing," Tech. Rep., 2014.

[19] H. Shirayanagi, H. Yamada, and K. Kono, "Honeyguide: A VM migration-aware network topology for saving energy consumption in data center networks," in Computers and Communications (ISCC), IEEE Symposium on, July 2012, pp. 460–467.

[20] S.-H. Wang, P.-W. Huang, C.-P. Wen, and L.-C. Wang, "EQVMP: Energy-efficient and qos-aware virtual machine placement for software defined datacenter networks," in Information Networking (ICOIN), International Conference on, Feb 2014, pp. 220–225.

[21] B. Puype et al., "Multilayer traffic engineering for energy efficiency," Photonic Netw. Commun., vol. 21, no. 2, Apr. 2011, pp. 127–140.

[22] W. Jiang, "Scalable ternary content addressable memory implementation using fpgas," in Architectures for Networking and Communications Systems (ANCS), ACM/IEEE Symposium on, Oct 2013, pp. 71–82.

 ISBN: 978-1-61208-398-8

On Multi-controller Placement Optimization in Software Defined Networking - based WANs

Eugen Borcoci, Radu Badea, Serban Georgica Obreja, Marius Vochin
University POLITEHNICA of Bucharest - UPB
Bucharest, Romania
eugen.borcoci@elcom.pub.ro, radu.badea.@elcom.pub.ro, serban@radio.pub.ro, mvochin@elcom.pub.ro

***Abstract* — Multi-controller implementation of the Software Defined Networking (SDN) control plane for large networks environment can solve the scalability and reliability issues introduced by the centralized logical control principle of the SDN. However, there are still open research topics related to controllers placement, static or dynamic assignment of the network forwarding nodes to controllers, especially when network nodes/links and/or controllers failures appear or some constraints are imposed. This paper contains an analytical view of some solutions proposed in the literature followed by a work in progress, on multi-criteria optimization methods applicable to the controller placement problem.**

Keywords* — *Software Defined Networking; Distributed Control Plane; Controller placement; Reliability; Multi-criteria optimizations.

I. INTRODUCTION

The recently proposed Software Defined Networking (SDN) technologies offer significant advantages for cloud data centres and also for Service Provider Wide Area Networks (WAN)..The *basic principles* of the SDN architecture are [1][2] [3]: *decoupling of the control and forwarding* (data) planes; *logically centralized control*; exposure of *abstract network resources and state to external applications.* Thus SDN offers important advantages of independency of the control software w.r.t. forwarding boxes implementations offered by different vendors. Higher network programmability is also a consequence of the above principles.

This paper considers the case when SDN–type of control is applied in a WAN, owned by an operator and/or a Service Provider (SP).

However, the control-data plane separation can generate performance limitations and also reliability issues of the SDN controlled network [4][5] (note that in the subsequent text, by "controller" it is understood a geographically distinct controller location):

(a) The forwarder nodes (called subsequently "forwarders" or simply "nodes") must be continuously controlled, in a proactive or reactive way. The forwarders have to ask their master controllers and then be instructed by them, how to process various new flows arriving to them (by filling appropriately the *flow tables* [1]). The control communication overhead (and its inherent delay), between several forwarders and a single controller, can significantly increase the response time of the overall system. This happens when the controller has a limited processing capacity [4], w.r.t the number of flow queries or the number of forwarders assigned to a controller is too high.

(b) The SDN control plane computes a single logical view upon the network; to this aim the controllers must inter-communicate and update/synchronize their data bases, in order to support the construction and continuously updating of unique vision upon the network [6][7][8]. A frequent solution for inter-controller communication is to create an overlay network linking the controllers on top of the same infrastructure used by the data plane flows [9].

(c) Asynchronous events such as controller failures or network disconnections between the control and data planes may also lead to packet loss and performance degradation [4][10]. Suppose that some forwarders are still alive (i.e., they can continue to forward the traffic flows conforming their current flow table content). However, if they cannot communicate with some controller, they will have no knowledge on how to process the newly arrived flows.

There is a need to optimally place the controllers. This can be done by attempting to solve (a), (b), and (c). This is a multi-criteria optimization problem and it was recognized as an NP-hard one [10]. Consequently, different solutions have been proposed targeting performance, (problem (a), (b)), and performance plus reliability (problem (c)).

This paper contains an analysis of some solutions for (a), (b), (c) and then proposes a preliminary contribution on how *multi-criteria optimization algorithms* can be applicable to the controller placement problem. The target here is *not to develop specific algorithms* dedicated to find an optimum solution for *a given criterion* (several studies did that) but to *achieve an overall controller placement optimization*, by applying *multi-criteria decision algorithms (MCDA)* [11][12]. The input of MCDA is a set of candidates (here an instance of controller placement is called a *candidate solution*).

The paper is organized as follows. Section II is an overview of related work. Section III outlines several metrics and algorithms used in optimizations and present some of their limitations. Section IV develops the framework for MCDA usage as a tool for final selection of the control placement solution. Section V presents conclusions and future work.

 ISBN: 978-1-61208-398-8

II. Related Work on SDN Controller Placement

This section is a short overview on some previously published work on controller placement in SDN-managed WANs. The basic problem to be answered (in an optimum way) is *how many SDN controllers are needed* in a given network (topology and some metrics are defined) and *where they should be placed* in the network, as to provide enough performance (e.g., low delay for controller-forwarder communications) and robustly preserve the performance level when failures occur. Intuitively, it can be seen that some trade-off will be necessary.

In WANs having significant path delays the controller placement determines the control plane convergence time, i.e., affects the controllers' response to real-time events sensed by the forwarders, or, in case of proactive actions, how fast can the controllers push (in advance) the required actions to forwarding nodes.

Actually, it has been shown in [10][13] that such a problem is theoretically not new. If *latency* is taken as a metric, the problem is similar to the known one, as *facility or warehouse location problem*, solved, e.g. by using Mixed Integer Linear Program (MILP) tools.

The Heller et al. early work [10] motivates the controller placement problem and then quantifies the placement impact on real topologies like Internet2 [4] and different cases taken from Internet Topology Zoo [15]. Actually, the main goal was not to find optimal minimum-latency placements (generally, such a problem has been previously solved)– but to present an initial analysis of a fundamental design problem, still open for further study. It has been shown that it is possible to find optimal solutions for realistic network instances, in failure-free scenarios, by analyzing the entire solution space, with off-line computations. This work also emphasized the fact (apparently surprising) that in most topologies, one single controller is enough to fulfill existing reaction-time requirements. However, resiliency aspects have not been considered in the above study.

Several works [9][13][16][17][18] have observed that resilience is important in the context of SDN and especially if Network Function Virtualization (NFV) is wanted. Some resiliency-related issues have been considered in [13]:

(1) *Controller failures*: in case of a primary controller failure, it should be possible to reassign all its previously controlled nodes to their second closest controllers, by using a backup assignment or signaling based on normal shortest path routing. Extreme case scenarios have been also considered, e.g., if at least one controller is still reachable, all nodes should keep functioning by communicating with it.

(2) *Network Disruption*: the failure of network links/ nodes, may appear, altering the topology. The routing paths (and their latencies) will change; some reassignment of nodes to other controllers is needed. In the worst case, some parts of the network can be completely cut off, having no access to controllers. Such nodes can still forward traffic, but they cannot anymore request or receive new instructions.

(3) *Controller overload* (load imbalance): shortest path–based assignment of the forwarders to controllers is natural. However one should avoid that one controller might have too many nodes to manage, otherwise its average response time will increase. Therefore, a well-balanced assignment of nodes to the different controllers is needed.

(4) *Inter-Controller Latency*: SDN concepts ask for a centralized logic view of the network, therefore inter—controller communications are necessary to synchronize their data bases. No matter if a single flat level of controllers (e.g., like in Onix [7]) or a hierarchical topology (e.g., like in Kandoo [8]) of controllers is used, it is clear that inter-controller latency should minimized. Therefore, an optimized controller placement should meet this requirement.

The works [9][17] present a metric to characterize the reliability of SDN control networks. Several placement algorithms are developed and applied to some real topologies, claiming to improve the reliability of SDN control, but still keep acceptable latencies. The controller instances are chosen such that the chance of connectivity loss is minimized; connections are defined according to the shortest path between controllers and forwarding devices.

The work [18] identifies several limitations of previous studies: (1) forwarder-controller connectivity is modelled using single paths, yet in practice multiple concurrent connections may be available; (2) peaks in the arrival of new flows are considered to be only handled on-demand, assuming that the network itself can sustain high request rates; (3) failover mechanisms require predefined information, which, in turn, has been overlooked. The paper proposes the *Survivor*, a controller placement strategy that explicitly considers path diversity, controller capacity awareness, and failover mechanisms at network design. Specific contributions consist in: significant reduction of the connectivity loss by exploring the path diversity (i.e., connectivity-awareness) which is shown to reduces the probability of connectivity loss in around 66% for single link failures; considering capacity-awareness proactively, while previous work handled requests churn on demand (it is shown that capacity planning is essential to avoid controller overload, especially during failover); smarter recovery mechanisms by proposing heuristics for defining a list of backup controllers (a methodology for composing such lists is developed; as a result, the converging state of the network can improve significantly, depending on the selected heuristic).

III. Metrics and Algorithms- Summary

This section summarizes some typical metrics and objectives of the optimization algorithms for controller placement. The overall goal is to optimize the Control Plane performance. Note that, given the problem complexity, the set of metrics and algorithms discussed below is not representing an exhaustive view. Considering a particular metric (criterion) an optimization algorithm can be applied, [9][10][13][18]. The goal of this paper is not to discuss details of such particular algorithms (but searching a global optimization). We only outline here their objectives. Some limitations are emphasized for particular cases.

 ISBN: 978-1-61208-398-8

A. *Performance-only related metrics (failure-free scenarios)*

The network is represented by an undirected graph G(V, E) where V is the set of nodes, n=|V| is the number of nodes and E is the set of edges. The edges weights represent an additive metric (e.g., *propagation latency* [10]). It is assumed that controller locations are the same as some of the network forwarding nodes.

A simple metric is d(v, c): *shortest path* distance from a forwarder node v∈V to a controller c∈V. In [10], two kinds of latencies are defined, for a particular placement C_i of controllers, where $C_i \subseteq V$ and $|C_i| \leq |V|$. The number of controllers is limited to $|C_i| = k$ for any particular placement C_i. The set of all possible placements is denoted by $C = \{C_1, C_2, \ldots.\}$. One can define, for a given placement C_i :

Average_latency:

$$L_{avg}(C_i) = \frac{1}{n} \sum_{v \in V} \min_{c \in Ci} d(v,c) \quad (1)$$

Worst_case_latency :

$$L_{wc} = \max_{v \in V} \min_{c \in C_i} d(v,c) \quad (2)$$

The optimization algorithm should find a particular placement C_{opt}, where *either average latency or the worst case latency is minimum.* Figure 1 shows a simple example of a network having six nodes. Two controllers $\{c_x, c_y\}$ can be placed in any location of the six nodes, e.g. in $\{v_5, v_6\}$. This placement instance is denoted by C_1. On the graph are marked the distances between different nodes (overlay paths)

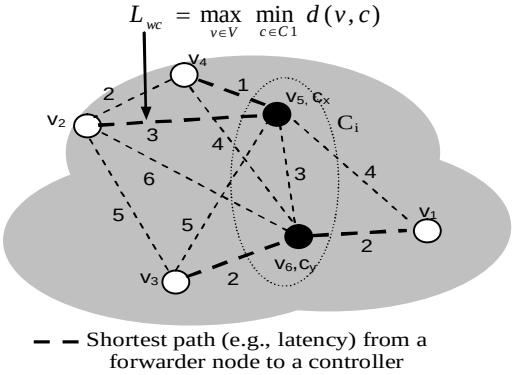


Figure 1. Simple network example of controller placement:v= forwarder node; c = controller; C_1= { [c_x_in_v_5 (v_5, v_2, v_4)], [c_y_in_v_6(v_6, v_1, v_3)]}

Some limitations of this optimization process are:

- No reliability awareness: the metrics are simply distances, which in the simplest case, are static.
- There is no upper limit on the number of v nodes assigned to a controller; too many forwarders to be controlled can exist, especially in large networks.

Other metric possible to be considered in failure-free case is *Maximum cover* [10][19]. The algorithm should find a controllers placement as to *maximize the number of nodes within a latency bound*; i.e., to find a placement of k controllers such that they cover a maximum number of forwarder nodes, while each forwarder must have a limited latency bound to its controller.

All metrics and algorithms described above do not take into account the inter-controller connectivity, so their associated optimizations as being partial..

B. *Reliability aware metrics*

Several studies consider more realistic scenarios in which controller failure or network links/nodes failure might exist. The optimization process aims now to find trade-offs (related to failure-free scenarios in order to assure still a convenient behavior of the overall system in failure cases.

(1) Controller failures (cf): the work [13] observes that node-to-controller mapping changes in case of controller outages. So, a realistic latency-based metric should consider both the distance to the (primary) controller and the distance to the other (backup) controllers. For a placement of a total number of k controllers, in [13] the failures are modelled by constructing a set C of scenarios, including all possible combinations of faulty controller number, from 0 of up to k - 1. The resulting maximum latency will be:

Worst_case_latency_cf :

$$L_{wc-cf} = \max_{v \in V} \max_{C_i \in C} \min_{c \in C_i} d(v,c) \quad (3)$$

The optimization algorithm should find a placement which *minimizes the expression* (3).

Commenting the placement results based on the metric (1) or (2) to (3), one can observe that in failure-free case the optimization algorithm tends to rather equally spread the controllers in the network, among the forwarders nodes. When minimization of expression (3) (and considering worst case failure) controllers tend to be placed the centre of the network. Thus, even if all except for one controller fail, the latencies are still satisfactory (numeric examples are given in [13]). However, one can criticize such an approach, if applied to large networks; the scenario supposed by the expression (3) is very pessimistic; rather a large network could be split in some regions/areas, each served by a primary controller; then some lists of possible backup controllers can be constructed for each area, as in [18].

The conclusion is that a trade-off exists, between the placements optimized for the failure free case and those including controller failure. It is a matter of operator policies to assign weights to different criteria before deciding, based on multiple criteria, the final selection of placement solution.

(2) Nodes/links failures (Nlf):

Links or nodes failures result in network disruption; some forwarders could have no more access to any controller. Therefore an optimization objective could be to find a controller placement which minimizes the number of nodes possible to enter into controller-less situations, in various scenarios of link/node failures. A realistic assumption is to limit the number simultaneous failures at only a few (e.g., two [13]). If more than two arbitrary link/node failures happen simultaneously, then the topology can be totally

 ISBN: 978-1-61208-398-8

disconnected and optimization of controller placement would not be any more useful.

For any given placement C_i of the controllers, an additive integer value metric *Nlf(Ci)* could be defined, as below:

- consider a failure scenario denoted by f_k, with $f_k \in F$, where F is the set of all network failure scenarios (in an instance scenario at most two link/nodes are down);
- initialize $Nlf_k(C_i) = 0$; then for each node $v \in V$, add one to $Nlf_k(C_i)$ if the node v has no path to any controller $c \in C_i$ and add zero otherwise;
- compute the maximum value (i.e., consider the worst failure scenario). We get:

$$Nlf(C_i) = \max Nlf_k(C_i) \quad (4)$$

where *k* covers all scenarios of F.

The *optimization algorithm* should find that *placement which minimizes (4).* It is naturally expected that increasing the number of controllers, will decrease the Nlf value. *We also observe that the optimum solution based on the metric (4) could be very different from those provided by the algorithms using the metrics (1) or (2).*

(3) Load balancing for controllers

A well designed system would require roughly equal load on all controllers, i.e., a good balance of the node-to-controller distribution. A metric can be defined to measure the degree of imbalance $Ib(C_i)$ of a given placement C_i as the *difference between the maximum and minimum number of nodes assigned to a controller.* If the failure scenarios set S is considered, then the worst case should evaluate the maximum imbalance as:

$$Ib(C_i) = \max_{s \in S} \{ \max_{c \in C_i} n_c^s - \min_{c \in C_i} n_c^s \} \quad (5)$$

where n_c^s is the number of forwarder nodes assigned to a controller c. Equation (5) takes into account that in case of failures the forwarders can be reassigned to other controllers than the primary ones and therefore, the load of those controllers will increase. An *optimization algorithm* should find that *placement which minimizes the expression (5).*

(4) Multiple-path connectivity metrics

One can exploit the possible multiple paths between a forwarder node and a controller [18], hoping to reduce the frequency of controller-less events, in cases of failures of nodes/links. The goal in this case is to maximize connectivity between forwarding nodes and controllers instances. The metric is :

$$M(C_i) = \frac{1}{|V|} \sum_{c \in C_i} \sum_{v \in V} ndp(v,c) \quad (6)$$

In (6), *ndp(v,c)* is the *number of disjoint paths* between a node v and a controller c, for an instance placement C_i. An *optimization algorithm should find the placement* C_{opt} *which maximizes* $M(C_i)$.

C. Inter-controller latency (Icl)

The inter-controller latency has impact on the response time of the inter-controller mutual updating. For a given placement C_i, the *Icl* can be given by the maximum latency between two controllers:

$$Icl(C_i) = \max d(c_k, c_n) \quad (7)$$

Minimizing (7) will lead to a placement with controllers close to each other. However this can increase the forwarder-controller distance (latency) given by (1) and (2). Therefore a trade-off is necessary, *thus justifying the necessity to apply some multi-criteria optimization algorithms, e.g., like Pareto frontier - based ones.*

D. Constraints

Apart from defining the metrics, the controller placement problem can be subject to different constraints. For instance, in [18], the input data for the optimal controller placement algorithm consists in the graph G(V,E) information, set of possible controller instances C, request demand of a network device, each controller capacity, and a backup capacity for each controller. Integer Linear Programming (ILP) –based algorithm is applied; here the constraints can be split into three classes: placement-related, capacity related and connectivity-related. In general other limits can be defined, e,g., on maximum admissible latency, ratio number controller/trivial nodes, regions pre-defined for controllers, etc. They should be included in the respective algorithms.

IV. MULTI-CITERIA OPTIMIZATION ALGORITHM

The sections II and III have shown that several criteria of optimum can be envisaged when selecting the best controller placement in a WAN. While particular metrics and optimization algorithms can be applied (see section III), we note that some criteria lead to partial contradictory placement solutions. What approach can be adopted? The answer can be given by adopting a multi-objective optimization based on *Multi-Criteria Decision Algorithms (MCDA).* The good property of MCDA is that it allows selection of a trade-off solution, based on several criteria. Note that partially such an approach has been already applied in [13] for some combinations of the metrics defined there (e.g., max. latency and controller load imbalance for failure-free and respectively failure use cases).

A. Reference level MCDA

We propose to apply MCDA, as a general way to optimize the controller placement, while considering not only a single metric but an arbitrary number of them.

The multi-objective optimization problem [11][12] is, to minimize $\{f_1(x), f_2(x), \ldots, f_m(x)\}$, where $x \in S$ (set of feasible solutions), $S \subset R^n$. The *decision vector* is $x = (x_1, x_2, \ldots, x_n)^T$. There are ($m \geq 2$) possibly conflicting objective functions $f_i : R^n \rightarrow R$, i= 1, ..m, and *we would want to minimize them simultaneously (if possible).* In controller placement problem we might have indeed some

partially conflicting objectives (e.g., to minimize the inter-controller latency and the forwarder-controller latency).

One can define *Objective vectors* = images of decision vectors. The objective (function) values are given by $z = f(x) = (f_1(x), f_2(x), \ldots, f_m(x))^T$. We denote as feasible objective region W = f (S) = image of S in the objective space.

Objective vectors are optimal if none of their components can be improved without deterioration to at least one of the other components.

A decision vector x_ ∈ S is named Pareto optimal [11] if there does not exist another x ∈ S such that $f_i(x) \leq f_i(x\_)$ for all i = 1, . . . , k and $f_j(x) < f_j(x\_)$ for at least one index j.

We adopt here the MCDA variant called *reference level decision algorithm* [12]. It has the advantage to allow selection of the optimal solution while considering normalized values of different criteria (metrics).

We use a simplified notation:

- identify the solutions directly by their images in the objectives space R^m ,
- decision parameters/variables are: v_i , i = 1, ..m, with $\forall i, v_i \geq 0$,
- image of a candidate solution is $Sl_s=(v_{s1},v_{s2},...,v_{sm})$, represented as a point in R^m ,

S = number of candidate solutions.

Note that the value ranges of decision variables may be bounded by given constrains. The optimization process consists in selecting a solution satisfying a given objective function and conforming a particular metric.

The basic *reference level algorithm* defines two reference parameters:

- r_i =*reservation level*=the upper limit for a decision variable, which the solution should not cross;
- a_i=*aspiration level*=the lower bound beyond which the reference parameters are seen as similar.

Without loss of generality one may apply the definitions of [12], where for each decision variable v_i there are defined r_i *and* a_i , by computing among all solutions s = 1, 2, ..S:

$$\begin{aligned} r_i &= \max[v_{is}], s = 1, 2, ..S \\ a_i &= \min\ [v_{is}], s = 1, 2, ..S \end{aligned} \quad (8)$$

In [12], modifications of the decision variables are proposed: *replace each variable* with *distance from it to the reservation level*: $v_i \rightarrow r_i - v_i$; (increasing v_i will decrease the distance); normalization is also introduced to get non-dimensional values, which can be numerically compared. For each variable v_{si}, a ratio is computed:

$$v_{si}' = (r_i - v_{si})/(r_i - a_i), \quad \forall s, i \quad (9)$$

The factor $1/(r_i-a_i)$ - plays also the role of a weight. The variable having high dispersion of values (max – min) will have lower weights, and so, greater chances to determine the minimum in the next relation (10). In other words, less preference is given to those variables having close values.

The basic algorithm steps are:

Step 0. Compute the matrix $M\{v_{si}'\}$, s=1…S, i=1…m

Step 1. Compute for each candidate solution *s*, the minimum among all its normalized variables v_{si}':

$$\min_s = \min\{v_{si}'\};\ i=1...m \quad (10)$$

Step 2. Make selection among solutions by computing:

$$v_{opt} = \max\ \{\ \min_s\ \}, s=1, ..S \quad (11)$$

This v_{opt} is the optimum solution, i.e. it selects the best value among those produced by the Step 1.

B. MCDA- Controller placement optimization

In this section, we *apply the reference level algorithm to the controller placement problem*. However, we modify the basic algorithm to be better adapted to controller placement problem, due to following remarks:

(1) The step 2 *compares values coming from different types of parameters/metrics* (e.g., max. latency, load imbalance, etc.) having different nature and being independent or dependent on each other. The normalization still allows them to be compared in the *max{ }* formula. *This is an inherent property of the basic algorithm.*

(2) However, the network provider might want to apply some policies when deciding the controller placement. Some decision variables (or metrics) could be more important than others. In some cases, the performance is more important, in others high resilience is the major objective.

A simple modification of the algorithm can support a variety of provider policies. We propose a modified formula:

$$v_{si}' = w_i(r_i - v_{si})/(r_i - a_i) \quad (12)$$

where the factor $w_i \in (0,1]$ represents a weight (priority) that can be established from network provider policy considerations, and can significantly influence the final selection.

The controller placement problem solving (given the graph, link costs/capacities, constraints, number of controllers desired, etc.) is composed of two macro-steps:

(1) *Macro-step1*: Identify the parameters of interest, and compute the values of the metrics for all possible controller placements, using specialized algorithms and metrics (1)-(7).

This procedure could be (depending on network size) time consuming and therefore performed off-line [10].

(2) *Macro-step2*: MCDA

- define reservation and aspiration levels for each decision variable;
- eliminate those candidates having parameter values out of range defined by the reservation level;
- define appropriate weights (see formula (9')) for different decision variables- depending on the high level policies applied by the operator;
- compute the normalized variables (formula (12))
- run the Step 0, 1 and 2 of the MCDA algorithm (formulas (10) and (11)).

 ISBN: 978-1-61208-398-8

The decision variables can be among those of Section III i.e.: *Average(1)* or worst(2) case latency (failure-free case); *Worst_case_latency_cf(3) Nodes/links failures (Nlf)(4); Controller Load imbalance(5); Multi-path connectivity metric(6); Inter-controller latency(7).*

For a particular problem, a selection of relevant variables should be done. E.g., in high reliable environment one could consider only failure free metrics.

C. Numerical example – MCDA optimization

Given the limited paper space, a simple but relevant example is exposed to illustrate the MCDA power, based on the Figure 1 network. Suppose that for this network the metrics of interest and decision variables are (see Section III) onl: d1:*Average latency* (1), d2: *worst latency (2)* (failure-free case); d3: *Inter-controller latency(7).* The reference levels are defined as in formula (8) and we propose: r_1=3, a_1=0;): r_2=6, a_2=0;): r_3=6, a_3=0.

Several placement samples can be considered:

C_1= { [c_x_in_v_5 (v_5, v_2, v_4)], [c_y_in_v_6(v_6, v_1, v_3)]}
C_2= { [c_x_in_v_5 (v_5, v_1, v_2, v_4)], [c_y_in_v_3(v_3, v_6)]}
C_3= { [c_x_in_v_3 (v_3,v_2)], [c_y_in_v_6(v_6, v_1, v_4, v_5,)]}
C_4= { [c_x_in_v_4 (v_4,v_2, v_5)], [c_y_in_v_6(v_6, v_1, v_3,)]}

1. MCDA with equal priorities for d1=1, d2=1, d3=1, The values of the metrics are computed using equations (1), (2) and respectively (7) for each placement: C_1, ..C4.

A matrix M(3x4) is computed using the formulas (9). MCDA is applied by using formulas (10), (11). The final result is : C_1 *= the best placement.* Looking at Figure 1, we indeed can see that this placement is a good trade-off between node-controller latency and inter-controller latency.

1. MCDA with priorities for i.e. d1=1, d2=0.5, d3=1, i.e., the worst case latency d2 has highest priority. After re-computing the matrix M and applying MCDA equations (1), (11), we find C_4*= the best placement.* Indeed we see in Figure 1 that worst case latency (node-controller) is minimized, however the inter-controller latency is higher than in C_1.

These examples proved how different provider policies can bias the algorithm.

V. CONCLUSIONS

This paper presented a work (in progress) study on using multi-criteria decision algorithms (MCDA for final selection among several controller placements solutions in WAN SDN, while considering several weighted criteria.

The method proposed is generic enough to be applied in various scenarios (including failure-free assumption ones or reliability aware), given that it achieves an overall optimization, based on multiple metrics supported by the reference model MCDA. Different network/service provider biases can be introduced in the selection process, by assigning policy-related weights to the decision variables.

Future work will be done to apply the method proposed to large networks - real life case studies (e.g. from Internet Topology zoo, [15]) and comparing the quality of trade-offs when defining different weights to decision variables.

REFERENCES

[1] B. N. Astuto, M. Mendonca, X. N. Nguyen, K. Obraczka, and T. Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks", Communications Surveys and Tutorials, IEEE Communications Society, (IEEE), 2014, 16 (3), pp. 1617 – 1634.

[2] "Software-Defined Networking: The New Norm for Networks" ONF White Paper 04.2012; retrieved: 02.2015 https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf.

[3] "SDN: The Service Provider Perspective", Ericsson Review, 21.02.2013. retrieved: 02.2015 http://www.ericsson.com/res/thecompany/docs/publications/ericsson_review/2013/er-software-defined-networking.pdf.

[4] S. H. Yeganeh, A. Tootoonchian and Y. Ganjali, "On Scalability of Software-Defined Networking", IEEE Comm. Magazine, February 2013, pp. 16-141,.

[5] M. Jarschel, F. Lehrieder, Z. Magyari, and R. Pries, "A Flexible OpenFlow-Controller Benchmark," in European Workshop on Software Defined Networks (EWSDN), Darmstadt, Germany, October 2012.

[6] A. Tootoonchian and Y. Ganjali, "Hyperflow: a distributed control plane for openflow" in Proc. INM/WREN, 2010.

[7] T. Koponen, M. Casado, N. Gude, J. Stribling, L. Poutievski, M. Zhu, R. Ramanathan, Y. Iwata, H. Inoue, T. Hama, and S. Shenker, "Onix: a distributed control platform for large-scale production networks," in Proc. OSDI, 2010.

[8] S. H. Yeganeh and Y. Ganjali, "Kandoo: A Framework for Efficient and Scalable Offloading of Control Applications," Proc. HotSDN '12 Wksp., 2012.

[9] H. Yan-nan, W. Wen-dong, G. Xiang-yang, Q. Xi-rong, C. Shi-duan, "On the placement of controllers in software-defined networks", ELSEVIER, Science Direct, vol. 19, Suppl.2, October 2012, pp. 92–97, http://www.sciencedirect.com/science/article/pii/S100588851160438X.

[10] B. Heller, R. Sherwood, and N. McKeown, "The controller placement problem," in Proc. HotSDN, 2012, pp. 7–12.

[11] J. Figueira, S. Greco, and M. Ehrgott, "Multiple CriteDecision Analysis: state of the art surveys", Kluwer Academic Publishers, 2005.

[12] A. P. Wierzbicki, "The use of reference objectives in multiobjective optimization". Lecture Notes in Economics and Mathematical Systems, vol. 177. Springer-Verlag, pp. 468–486.

[13] D. Hock, M. Hartmann, S. Gebert, M. Jarschel, T. Zinner, and P. Tran-Gia, "Pareto-Optimal Resilient Controller Placement in SDN-based Core Networks," in ITC , Shanghai, China, 2013.

[14] Internet2 open science, scholarship and services exchange. http://www.internet2.edu/network/ose/.

[15] S. Knight, H. X. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet Topology Zoo," IEEE JSAC, vol. 29, no. 9, 2011.

[16] Y. Zhang, N. Beheshti, and M. Tatipamula, "On Resilience of Split-Architecture Networks," in GLOBECOM 2011, 2011.

[17] Y. Hu, W. Wendong, X. Gong, X. Que, and C. Shiduan, "Reliability aware controller placement for software-defined networks," in Proc. IM. IEEE, 2013, pp. 672–675.

[18] L. Muller, R. Oliveira, M. Luizelli, L. Gaspary, M. Barcellos, "Survivor: an Enhanced Controller Placement Strategy for Improving SDN Survivability", IEEE Global Comm. Conference (GLOBECOM); 12/2014.

[19] D. Hochba "Approximation algorithms for np-hard problems", ACM SIGACT News, 28(2), 1997, pp. 40–52.

 ISBN: 978-1-61208-398-8